DOCUMENT RESUME

ED 356 260 TM 019 685

AUTHOR Holburn, P. T.
TITLE Test Bias in the Intermediate Mental Alertness, Mechanical Comprehension, Blox and High Level Figure Classification Tests. An NTB/HSRC Report.
INSTITUTION Human Sciences Research Council, Pretoria (South Africa).
REPORT NO C/PERS-453; ISBN-0-7969-1254-8
PUB DATE 92
NOTE 84p.; Paper includes English and Afrikaans abstracts.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Adults; *Apprenticeships; Blacks; Comparative Testing; Correlation; *Culture Fair Tests; Foreign Countries; *Job Applicants; Occupational Tests; *Personnel Selection; *Racial Differences; Social Bias; *Test Bias; Test Reliability; Test Use; Whites
IDENTIFIERS Asians; Blox Test; High Level Figure Classification Test; Intermediate Mental Alertness Test; Mechanical Comprehension Test (Physics); *South Africa

ABSTRACT

Research is reported on four tests commonly used in South Africa to select apprentices, the Intermediate Mental Alertness Test, the High Level Figure Classification Test, the Blox Test, and the Mechanical Comprehension Test. Samples were as follows: (1) 206 Asian, 208 Black, 102 Coloured, and 99 White mostly male applicants for sugar industry positions; (2) Black, mostly male applicants for a development company; and (3) 14 Asian, 128 Black, 199 Coloured, and 74 White applicants for a motor company. The Mechanical Comprehension Test was found to have low internal consistency reliability figures for Asian, Black, and Coloured applicants, and it was not recommended for multicultural apprentice selection. Item bias detection procedures were applied to the other tests. Overall, it appears that the Intermediate Mental Alertness Test had the most bias, followed by the Blox Test. Very little bias emerged for the High Level Figure Classification Test. Despite the presence of bias, these tests are recommended for use in batteries for apprentices. Suggestions are given for dealing with test bias. Ten tables present study findings and one figure illustrates predictive bias. Three appendixes list item difficulty values, item-total correlations, and percentages not completing an item for the three tests. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Contract Report C/PERS 453

ED356260

TEST BIAS IN THE INTERMEDIATE MENTAL ALERTNESS, MECHANICAL COMPREHENSION, BLOX AND HIGH LEVEL FIGURE CLASSIFICATION TESTS

an NTB/HSRC report

P T Holburn

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. G. GARBERS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM019685

BEST COPY AVAILABLE

Contract Report C/PERS 453

TEST BIAS IN THE INTERMEDIATE MENTAL ALERTNESS, MECHANICAL COMPREHENSION, BLOX AND HIGH LEVEL FIGURE CLASSIFICATION TESTS

An NTB/HSRC report

P T HOLBURN

Pretoria
Human Sciences Research Council
1992

P T Holburn MSc. Senior Researcher

Group: Human Resources
General Manager: Dr K F Mauer

ISBN 0 7969 1254 8

© Human Sciences Research Council, 1992

Printed and published by the HSRC
134 Pretorius Street, Pretoria.

Acknowledgements

Dr K F Mauer: General Manager, Human Resources

Mr R F Skawran: Manager, Human Development

Dr T R Taylor: Head, Cognitive and Personality Research Programme

National Training Board

KnowledgeTec, Johannesburg

Jane England: Human Development Group

Ricka Richter: Human Development Group

Special thanks are due to Dr T R Taylor, who made many useful comments and suggestions to previous drafts of this report.

ABSTRACT

In this document research examining test bias in the Intermediate Mental Alertness test, High Level Figure Classification test, Blox test and Mechanical Comprehension test for apprentice applicants is reported.

The Mechanical Comprehension test was found to have very low internal consistency reliability figures for the Asian, black and coloured applicants, and consequently it was recommended that this test not be used for multicultural apprentice selection.

Item bias detection procedures were applied to the Intermediate Mental Alertness, Blox and High Level Figure Classification tests. Seven of the items of the Intermediate Mental Alertness were biased against black applicants, one High Level Figure Classification test item was biased against the black group, while three items from the Blox test were biased against black applicants and three different items biased against coloured applicants.

Overall, with the predictive bias results included, it appeared that the Intermediate Mental Alertness test had the most bias, followed by the Blox test. Very little bias emerged for the High Level Figure Classification test.

Despite the presence of bias, it was recommended that the test battery for apprentices should include the Intermediate Mental Alertness test, the High Level Figure Classification test, and the Blox test, together with a new HSRC Mechanical test battery that is currently being completed. Different ways of dealing with the test bias were proposed in the report.

EKSERP

In hierdie dokument word verslag gedoen oor navorsing wat toetssydigheid in die Intermediêre Verstandelike Helderheidstoets, Hoëvlak Figuurindelingstoets, Bloxtoets en Meganiese Insigtoets ondersoek, soos dit betrekking het op vakleerlingaansoekers.

Daar is gevind dat die Meganiese Insigtoets baie lae interne konsekwentheid betroubaarheidstellings vir die Asiër, swart en gekleurde aansoekers het, en gevolglik is aanbeveel dat hierdie toets nie gebruik word vir die keuring van vakleerlinge in multikulturele situasies nie.

Prosedures vir die opsporing van itemsydigheid is toegepas op die Intermediêre Verstandelike Helderheids-, Blox en Hoëvlak Figuurklassifikasietoetse. Sewe van die items uit die Verstandelike Helderheidstoets was sydig teen swart aansoekers, een Hoëvlak Figuurindelingstoets-item was sydig teen die swart groep, terwyl drie items uit die Bloxtoets sydig was teen swart aansoekers en drie ander items sydig was teen gekleurde aansoekers.

Wat die resultate as 'n geheel betref, insluitend die voorspellende sydigheidsresultate, blyk dit dat die Intermediêre Verstandelike Helderheidstoets die sydigste is, gevolg deur die Bloxtoets. Die Hoëvlak Figuurindelingstoets het baie min sydigheid opgelewer.

Ten spyte van die teenwoordigheid van sydigheid, word aanbeveel dat die toetsbattery vir vakleerlinge die Intermediêre Verstandelike Helderheidstoets, die Hoëvlak Figuurindelingstoets and die Bloxtoets, asook die nuwe RGN Meganiese toetsbattery wat tans voltooi word, moet bevat. Verskillende maniere om vir die toetssydigheid te kompenseer is voorgestel in die verslag.

CONTENTS

	PAGE
1. INTRODUCTION	1
2. PROCEDURE USED TO DETECT TEST BIAS	5
2.1 Item bias	5
2.2 Predictive bias	7
2.3 General	9
3. DESCRIPTION OF THE SAMPLES AND TESTS	10
3.1 Description of the tests	10
3.1.1 The Intermediate Mental Alertness	10
3.1.2 High Level Figure Classification Test (A/129)	10
3.1.3 Blox (A/80)	11
3.1.4 Mechanical Comprehension (A/3/1)	11
3.2 Description of the samples	12
3.2.1 Sample 1	12
3.2.2 Sample 2	12
3.2.3 Sample 3	13
4. ITEM BIAS PROCEDURE	15
5. ITEM BIAS ANALYSES	17
5.1 General	17
5.2 Item bias results	18
5.2.1 Intermediate Mental Alertness	18
5.2.1.1 Means, standard deviations and KR20 statistics	18
5.2.1.2 Black-white comparisons - Item bias results	19
5.2.1.3 Coloured-white comparisons - Item bias results	20
5.2.1.4 Asian-white comparisons - Item bias results	21
5.2.1.5 Summary of findings	21
5.2.2 High Level Figure Classification test	24
5.2.2.1 Means, standard deviations and KR20 statistics	24
5.2.2.2 Black-white comparisons - Item bias results	24
5.2.2.3 Coloured-white comparisons - Item bias results	25
5.2.2.4 Asian-white comparisons - Item bias results	26
5.2.2.5 Summary of findings	26
5.2.3 Blox test	29
5.2.3.1 Means, standard deviations and KR20 statistics	29
5.2.3.2 Black-white comparisons - Item bias results	30
5.2.3.3 Coloured-white comparisons - Item bias results	30
5.2.3.4 Asian-white comparisons - Item bias results	30
5.2.3.5 Summary of findings	30
5.2.4 Mechanical Comprehension test	33
5.2.4.1 Means, standard deviations and KR20 statistics	33
5.2.4.2 General	33

6. PREDICTIVE BIAS ANALYSES 35
6.1 Introduction 35
6.2 Test information 37
6.3 Criterion data 37
6.3.1 Progress Evaluation Report (C1) 38
6.3.2 Efficiency Report (C2) 38
6.4 Description of the sample 38
6.5 Procedure 39
6.6 Predictive bias results for the High Level Figure Classification test 41
6.6.1 High Level Figure Classification test and C1 41
6.6.2 High Level Figure Classification test and C2 41
6.7 Predictive bias results for the Blox test 42
6.7.1 Blox test and C1 42
6.7.2 Blox test and C2 42
6.8 Summary 42
7. SUMMARY AND CONCLUSIONS 44
7.1 Item bias 44
7.1.1 Item bias - Intermediate Mental Alertness test 44
7.1.2 Item bias - High Level Figure Classification Test 46
7.1.3 Item bias - Blox 47
7.1.4 Item bias - Mechanical Comprehension test 47
7.1.5 Item bias - Conclusion 48
7.2 Predictive bias 50
7.3 Dealing with item and predictive bias 52
7.4 Conclusion and recommendations 57
REFERENCES 60
APPENDIX A 64
APPENDIX B 67
APPENDIX C 70

LIST OF TABLES

		PAGE
Table 3.1	Age and education of sample 1	12
Table 3.2	Age, education and home language of sample 3	13
Table 5.1	Means, standard deviations and KR20 statistics for the Intermediate Mental Alertness test	19
Table 5.2	Original and rescored means and standard deviations for the Intermediate Mental Alertness test	22
Table 5.3	Means, standard deviations and KR20 statistics for the High Level Figure Classification test	25
Table 5.4	Original and rescored means and standard deviations for the High Level Figure Classification test	27
Table 5.5	Means, standard deviations and KR20 statistics for the Blox test	29
Table 5.6	Original and rescored means and standard deviations for the Blox test	31
Table 5.7	Means, standard deviations and KR20 statistics for the Mechanical Comprehension test	34
Table 6.1	Predictive bias analyses undertaken	40

LIST OF FIGURES

PAGE

Figure 7.1 Illustration of predictive bias 52

1. INTRODUCTION

Many organisations in South Africa make use of psychometric tests to select personnel for various jobs within their companies. In the 1980s, however, several test users began to question the suitability of tests for all applicants. Consequently, in the past few years many employers have become concerned with the fairness of tests used for selection and placement, as well as for training and promotion purposes. There is also a growing concern from the side of the applicants who may perceive they have been unjustly excluded from jobs due to the use of biased tests.

Cross cultural research has revealed that different cultural groups obtain different average scores on psychological tests. This finding has led to suggestions that tests are culture-biased, and recommendations that psychologists concentrate on developing culture-free tests have been made. Some tests that were developed in response to this argument, and hence purported to be culture-free, include the Raven's Progressive Matrices and Goodenough Draw-A-Man Test (Anastasi, 1985; Kaplan, 1985). However, it was subsequently realised that culture permeates all aspects of one's thinking and feeling and therefore any attempts to develop a culture-free test are elusive.

The next issue that arose in response to the accusation of cultural bias in tests was that of culture-fair tests. If culture-free tests could not be developed, perhaps culture-fair tests were the answer. Such tests generally comprised non-verbal figural type items, the content of which was assumed to be equally familiar to all cultures. Nevertheless, such material is also culturally dependent. Different cultures do not make use of nonverbal and figural material in the same way and to the same extent. Different cultures emphasise the development and use of different thought processes and skills, and the manner in which figural material is manipulated in different cultures varies,

even if the material is equally familiar to all cultural groups (which is extremely unlikely).

Despite general agreement that the development of culture-free and culture-fair tests is difficult, if not impossible, the debate surrounding culture and testing and its concomitant problems has remained. Tests have continued to be accused of being biased, and the use of tests is particularly problematic when a group of examinees comprise members of different cultural groups. During the latter half of this century the applicants for jobs and places in educational institutions have increasingly come from diverse backgrounds; hence, concern over test bias has increased. It has thus become imperative that test users and test developers give serious thought to evaluating tests for bias. To aid this process, and because of the seriousness of the issue, during the 1970s much literature was published in the educational and psychological journals outlining procedures for conducting test bias studies.

One of the first major studies to investigate test bias was conducted by Eells and colleagues in the 1950s. Their research focused on test bias amongst different socioeconomic groups (Eells et al., 1951). Although average test scores between different socioeconomic groups differ (persons of low socioeconomic status [SES] obtain lower scores on average than high SES individuals), test bias research has, by and large, tended to concentrate on ethnic groups rather than groups of different socioeconomic status.

Despite the obvious importance of conducting research in the area of test bias, a start has only recently been made in South Africa. Some studies investigating bias in psychometric tests have been conducted recently by the Human Sciences Research Council (HSRC) (Claassen, 1990; Claassen & Cudeck, 1985; Cudeck & Claassen, 1983; Owen, 1986; Owen, 1989). This research focused on tests used in the educational context, and although the findings from these studies are extremely useful, much more

needs to be done at the organisational level, particularly since at present most cross cultural selection using tests occurs in the workplace.

A report (Taylor, 1987) has been published explaining in detail the meaning of test bias and the various methods that can be used to evaluate tests for bias. Thus in this report we will not be concerned with a theoretical discussion of test bias. The purpose of this research was to empirically examine tests for bias and to report the results. Test bias was examined using the procedures outlined by Taylor (1987).

The groups that are considered when we compare test scores to ascertain if tests are fair are usually race groups. There are several reasons for this.

Firstly, because of legislation which has in the past prohibited certain groups from entering certain job categories, many of the tests currently used for selection and placement have been developed for, and standardised on, whites only. Because all jobs are now open to all and are increasingly being filled by applicants from all race groups, there is a need to ascertain the suitability of tests across races. At this stage, we know little of the comparability of test scores for different cultural groups in South Africa.

Secondly, the different race groups in South Africa have lived very separate lives. The use of a common test for all applicants relies on the assumption that the candidates have all been exposed to the same developmental opportunities. This is unlikely to be true of the different races in South Africa. Different race groups live in different areas, attend different schools under different education systems and speak different languages. These factors provide reasons for us to suspect that bias may be present in tests.

Currently most of the unskilled and semi-skilled jobs in South

Africa are held by blacks, whereas most of the professional and managerial jobs are held by whites. As more individuals from groups other than the white group take advantage of the opening up of jobs to all races, we can expect that for the jobs at the lower end of the skills spectrum there will be many applicants from all races. That is, one would expect apprentice applicants to be from all race groups. Thus it is important that we determine the cross-cultural comparability of test scores of apprentice applicants. The tests most frequently used to select apprentices have been identified in a survey (Holburn, 1989), and in the remainder of this report the results of bias analyses undertaken on these tests will be reported.

2. PROCEDURE USED TO DETECT TEST BIAS

Two types of bias associated with tests were examined in this report: Item bias and predictive bias. Predictive bias is discussed in more detail in chapter 6.

2.1 ITEM BIAS

An item that is more difficult for one group than another is not necessarily biased. Difficulty and bias are not the same thing. Item bias is concerned with the extent to which an item functions anomalously for different groups. An item is biased if it is more or less difficult for a group relative to the group's performance on the other items in a test. Taylor (1987) noted that there are two definitions of item bias, one unconditional and one conditional.

According to the **unconditional** definition of item bias an item is biased if, on that item, members of one group obtain an average score which differs from the average score of the other group by more or less than would be expected from the group's performance on the other test items. That is, an item is biased if it is relatively easier for one group than another. In ANOVA terms this represents an item X group interaction. Examples of unconditional item bias detection methods include ANOVA and the Transformed Item Difficulties method (TID).

The **conditional** definition of item bias specifies that an item is biased if the probability of a correct response differs for groups of the same ability level. This is called the conditional definition of item bias because bias is defined conditional upon ability level. Examples of conditional item bias detection methods include methods based on item response theory and contingency tables.

The set of items identified as biased by a particular item bias

detection method is likely to differ from the set of biased items produced by another method because each method is based on different theoretical approaches. One should, however, if the item bias detection methods have reasonable validity and reliability, expect a great deal of overlap between the items identified as biased by the different methods.

Taylor (1987) has reviewed the literature and evaluated the various item bias detection methods on theoretical and empirical grounds. On a theoretical and empirical level the three parameter item characteristic curve performs the best, but because large samples of at least 1000 subjects are required for each group, this item bias detection method should not be used for investigating bias in tests used in industry where the numbers of blacks are small at present.

Apart from the three parameter item characteristic curve method, methods based on contingency tables and the TID method have fared reasonably well in research aimed at evaluating item bias detection methods (Ironson & Subkoviak, 1979; Rudner et al., 1980; Shepard et al., 1985; Subkoviak et al., 1984). In particular, studies have found that methods based on contingency tables produce sets of biased items very similar to those of the three parameter item characteristic curve. From a theoretical point of view, one of the best methods for detecting biased items is the logit method, and several authors have pointed out that the use of an iterative item bias detection method is preferable to a non-iterative technique (Lord, 1977; Mellenbergh, 1982; Taylor, 1987; Van der Flier et al., 1984).

A strategy for detecting biased items has been proposed by Taylor (1987). This strategy entails the use of a multi-method multi-sample approach. Taylor (1987) recommended applying two item bias detection methods to several samples of examinees. He argued that one of the item bias detection methods should be based on the conditional definition and one on the unconditional definition, and that the samples of subjects should differ in some way, e.g.

come from different geographical areas. The set of items that is ultimately identified as biased will depend on the frequency with which the different methods identify items as biased from the different samples.

To date the TID method is probably the best of the unconditional methods for detecting biased items and is the unconditional method recommended by Taylor (1987). Because it is not feasible to use the three parameter item characteristic curve with the small samples typically obtained in industry, Taylor (1987) recommended that the conditional item bias detection method applied should be the iterative logit method. These methods are discussed in detail in Taylor (1987).

It must be borne in mind that no item bias detection method can detect pervasive bias. Pervasive bias refers to bias in a test which affects all of the items equally. Therefore if the analyses do not reveal any bias, one cannot conclude there is definitely no bias in the tests. It is also important to realise that item bias is a matter of degree. Reference to an item as biased or unbiased refers to whether the bias in an item is considered to be relatively minimal or substantial. Critical values based on statistical tests are used to determine cut-off points for demarcating an item as biased or unbiased, and clearly some subjectivity is involved in setting critical values.

2.2 PREDICTIVE BIAS

In practical situations tests are used to predict some criterion score, for example job performance or training course performance. The test scores of job applicants are entered into a regression equation and the expected criterion score calculated. The prediction of job performance from test scores is of major importance to the test user; hence, the examination of tests for predictive bias should be of great concern to employers.

Predictive bias exists if the regression equations used to predict criterion scores are not the same for different groups. Predictive bias is defined as follows: "A test is a biased predictor if there is a statistically significant difference between the major and minor groups in the slopes, or in the intercepts, or in the standard error of estimates of the regression lines of the two groups, when these regression parameters are derived from the estimated true scores of persons within each group" (Jensen, 1980, p 381-382). Therefore, in order to empirically investigate predictive bias it is necessary to compare the slopes, intercepts and standard error of estimates of the regression lines of the different groups. If there is a significant difference on any one of these three regression parameters the test is predictively biased for that particular criterion.

In prediction studies we are concerned with predicting performance on a particular criterion. Thus a test may be predictively biased with respect to one particular criterion but not another. This is why Taylor (1987) has argued that no single predictive bias study can determine if a test is predictively biased. Thus the evaluation of predictive bias is going to fall largely in the hands of the test user. However, it was felt that at least one predictive bias study should be presented in this report in order to demonstrate how one would conduct such a study, as well as to reveal findings of interest, even if of limited generalisability.

One of the biggest problems in conducting predictive validity research is the choice of a suitable criterion. Similarly, in a predictive bias study, careful consideration should be paid to the selection of a criterion. The criterion must be the same for all groups (criteria of training course results where different groups have attended different training institutions do not suffice) and should be unbiased.

2.3 GENERAL

In this report the focus was on item bias studies because this area is the responsibility of, and is most easily and adequately performed by, the test developer. A predictive bias study was also undertaken. However, a single study cannot provide definitive results on predictive bias and the onus is on the test user to assess whether a test is predictively biased for a particular use of the test.

3. DESCRIPTION OF THE SAMPLES AND TESTS

A report on apprentice selection in South Africa (Holburn, 1989) revealed that the tests most frequently used for apprentice selection (apart from the Department of Manpower Aptitude Test Battery) are four NIPR tests. These four tests are the Intermediate Mental Alertness (B/77), the High Level Figure Classification Test (A/129), the Blox test (A/80) and the Mechanical Comprehension test (A/3/1). These tests are briefly described below. However, in order to ensure confidentiality, the following discussion only contains basic information about the tests.

3.1 DESCRIPTION OF THE TESTS

3.1.1 *The Intermediate Mental Alertness*

This test forms part of the Intermediate Battery (B/77) (and therefore may be administered together with the remainder of the tests in the battery). However, the Intermediate Mental Alertness (Int MA) is often administered on its own.

The Intermediate Mental Alertness is a multiple choice test of 30 items, suitable for candidates with nine to twelve years of formal schooling. Verbal and numerical reasoning ability is measured by the Mental Alertness test, and items include alphabetic codes, alphabetic series, numeric series, similarities, verbal analogies and math word problems (Wilcocks, 1973).

3.1.2 *High Level Figure Classification test (A/129)*

The High Level Figure Classification test (HL FCT) measures nonverbal abstract conceptual reasoning ability (Werbeloff & Taylor, 1983). The NIPR has developed a standard level version for use with candidates who have less than ten years of formal

schooling. The high level version is appropriate for testing candidates with ten to twelve years of scholastic education, and hence is appropriate for the selection of apprentices and technicians.

The HL FCT was developed out of a need for a nonverbal test of general intelligence suitable for all cultural groups. Most of the tests that had been developed up until this time were for whites only, or else separate tests had been constructed for whites and blacks. Werbeloff and Taylor (1982) have contended that the HL FCT should provide a good index of an individual's general intellectual ability without the influence of verbal skills. Werbeloff and Taylor (1982) further argued that the test is suitable for Asians and blacks and reported that the HL FCT correlated well with criterion scores (technician test results) for black male technicians.

3.1.3 *Blox (A/80)*

The Blox test, formerly known as the Perceptual Battery, is a measure of the ability to recognise spatial arrangements from different orientations. This ability is considered to be a subset of spatial ability (Halstead & du Toit, 1983). The Blox test is suitable for examinees who have 9 to 12 years of formal schooling, and in the test administrators manual it is pointed out that this multiple choice test favours those who have had experience and training in technical drawing and mechanics.

3.1.4 *Mechanical Comprehension (A/3/1)*

The Mechanical Comprehension test is a multiple choice test designed to measure the ability to apply knowledge of the laws and principles of physics appropriately. The test is considered to be useful in the selection and classification of personnel in technical fields (Visser, 1978). This test is suitable for applicants who have 9 to 12 years of formal schooling. Different versions of this test are in use, but it is the revised 42 item

test that is most commonly applied and which is discussed in this report.

3.2 DESCRIPTION OF THE SAMPLES

In accordance with the bias detection strategy adopted and discussed earlier, numerous samples of data were required. Because this research formed part of a larger project investigating apprentice selection procedures, the samples comprised applicants for apprentice positions in industry. Three main samples of applicants were used in the item bias studies.

3.2.1 *Sample 1*

The first sample comprised Asian, black, coloured and white applicants for apprentice positions in the sugar industry in Natal. These applicants, the vast majority of which were male, completed the Intermediate Mental Alertness, High Level Figure Classification test, Blox test and Mechanical Comprehension test. In Table 3.1 the statistics describing the age and education of this sample are presented.

TABLE 3.1 AGE AND EDUCATION OF SAMPLE 1

	Year tested	Age range	Mean age	Years schooling	Post school *	N
Asians	87-88	16-37	20,55	10-12	34,47	206
Blacks	86-88	18-33	23,00	10-12	65,38	208
Coloureds	81-88	17-31	20,06	10-12	33,33	102
Whites	81-88	17-27	19,77	10-12	20,20	99

* Percentage with a technical college qualification e.g. N1, N2

3.2.2 *Sample 2*

A development company in Natal tested a group of applicants for apprentice positions. All the applicants were black and

completed the Intermediate Mental Alertness, High Level Figure Classification test, Blox test and the Mechanical Comprehension test.

The vast majority of this sample were male (98%) and spoke zulu at home. They ranged in age from 18 to 38 years, with an average age of 23 years. The examinees were tested between 1985 and 1988 and had completed between ten and twelve years of formal education.

3.2.3 *Sample 3*

The third sample on which the analyses in this report were based, were apprentice applicants for a large motor company in the Eastern Cape. The examinees, of all race groups and mostly male (over 90%), were tested towards the end of 1989. Each applicant completed the Intermediate Mental Alertness test, the High Level Figure Classification test and the Blox test. Table 3.2 contains the information summarising the age, educational and home language characteristics of the sample.

TABLE 3.2 AGE, EDUCATION AND HOME LANGUAGE OF SAMPLE 3

	Age range	Mean age	Years schooling	Post school *	Home lang	N
Asians	18-23	19,5	10-12	28,57	mostly English	14
Blacks	17-38	23,2	10-12	34,38	Xhosa	128
Coloureds	15-29	20,7	10-12	28,14	mostly Afrikaans	199
Whites	16-25	19,3	10-12	28,38	Afrik & English	74

* Percentage with a technical college qualification, e.g. N1, N2

The item bias analyses were conducted on the test scores of these three samples. If these tables are examined it can be seen that

there was a tendency for the black applicants to be somewhat older than the other race groups, with the white group generally being the youngest. With regard to post-school qualifications, more black than Asian, coloured or white applicants had been to technical college.

4. ITEM BIAS PROCEDURE

In order to conduct any bias study, a minimum of two groups are compared. With the available data in this project many different comparisons were possible for each test. With four defined groups, Asians, blacks, coloureds and whites, it would have been possible to compare whites and blacks, whites and coloureds, blacks and Asians, Asians and coloureds etc. However, in an attempt to deal with the data in a thorough yet manageable way, it was decided that because the tests were developed primarily for whites, they would be assumed to be unbiased for this group. Thus the only comparisons required would be white-Asian, white-black and white-coloured comparisons. (The reader should not infer that other comparisons are irrelevant or uninteresting, but rather that the time and space for dealing with the data is limited, rendering this the best approach.)

The item bias results are reported separately for each test. For each test, item bias analyses were performed on every possible unique white-Asian, white-black and white-coloured pair. The available samples permitted six black-white comparisons, four coloured-white comparisons and two Asian-white comparisons.

The TID and iterative logit item bias detection methods were applied to the data. Bias was determined in items for the TID analysis by calculating the perpendicular distance an item fell from the major axis line. If an item was 0,75 or more z score units away from the line the item was identified as biased. This cut-off point is the one that is usually adopted. The iterative logit method is associated with a statistical test (chi-square) and hence significance values can be determined. If an item had a probability value of 0,05 or less, the item was considered biased.

These two item bias detection methods were applied to all possible white-Asian, white-black and white-coloured pairs of

data and the number of times an item emerged as biased within any particular comparison noted. If an item was flagged as biased in 50% or more of the white-black comparisons, that item was identified as definitely biased. The same procedure was applied for the white-Asian and white-coloured comparisons. Thus an item of a test was considered to be definitely biased if it was biased in three or more of the black-white comparisons, or in two or more of the four coloured-white comparisons. Because only two Asian-white comparisons could be performed, an item was considered biased if it emerged as biased in either one or both of the comparisons.

Baseline comparisons were also undertaken. A baseline comparison is an item bias analysis between two samples from the same race group. If an item is biased in a comparison between two samples from the same race group (but possibly different industries or geographical areas) that item is biased at a baseline level. The rationale underlying baseline studies is that if an item is to be considered biased in a black-white comparison, the bias index has to be higher than the bias indices in black-black and white-white baseline studies.

In this research one black-black, one coloured-coloured and one white-white baseline comparison was undertaken between the sample 1 and sample 3 subjects. That is, an item bias analysis was performed between sample 1 and sample 3 whites, sample 1 and 3 blacks and sample 1 and 3 coloureds. In the item bias research reported in the next chapter, if biased items were found when two different races were compared and the same items were biased in the same race baseline studies, then that item was not reported as biased.

5. ITEM BIAS ANALYSES

5.1 GENERAL

When the results of the TID procedure were examined it was decided that this technique was not adequate for detecting bias in these specific samples. In the literature it has been noted that unconditional methods of item bias detection, such as the TID, do not work well when two groups which differ substantially in average test scores are compared (Ironson & Subkoviak, 1979; Linn et al., 1981). South African samples of blacks and whites usually have average scores that are fairly different; blacks usually have an average test score more than one standard deviation below that of whites. Under these circumstances, when an item bias detection technique based on the unconditional definition is used, bias and item discriminability often become confounded. Therefore the items that discriminate the most effectively between good and poor performers (i.e. are good test items in that the test is intended to differentiate the good from the weaker applicants) often emerge as biased items. When the item bias analyses were performed on the data, it was found that almost every item was flagged as biased in many of the studies using the TID. This result could have occurred because in almost all samples the black-white mean differences were very large.

It is possible to correct for large average score differences by constructing pseudogroups (Angoff, 1982; Shepard et al., 1984). Cases are deleted from the white or black samples until samples (pseudogroups) that have approximately the same average score on the test have been constructed. However, where large mean differences between groups exist this is not easily accomplished. Hence it was decided to use the results of the iterative logit method only. It is also necessary to point out that because South African samples tend to display sizeable black-white mean differences, caution should be exercised when interpreting the results of a TID item bias analysis.

The following bias findings are, therefore, based on the iterative logit method only. It is the author's opinion that item bias findings based on the iterative logit method alone are adequate, especially since several samples were involved. It was pointed out earlier that the iterative logit method is based on the conditional definition of item bias, which is from a theoretical and practical point of view superior to any unconditional item bias method (Mellenbergh, 1982; Taylor, 1987).

5.2 ITEM BIAS RESULTS

5.2 1 *INTERMEDIATE MENTAL ALERTNESS*

5.2.1.1 Means, standard deviations and KR20 statistics

The means, standard deviations and internal consistency reliability figures for the samples are listed in Table 5.1. The samples are as described in chapter 3.

In the South African research literature on testing it can be observed that the highest mean scores are generally those of whites, followed by Asians and coloureds, then blacks. The mean test scores in sample 1 are typical in this regard. Sample three was unusual in that the white mean score was below that of the coloured group. This reflects the nature of the sample, i.e. the job applicants, and should not be taken as representative of the population in general. Sample 3 was also unusual in that the standard deviation for the white group was much larger than the standard deviations of the other groups. However, in the other samples, the standard deviations were fairly similar.

With regard to internal consistency reliability, it can be seen that the reliability for Asians, coloureds and whites was acceptable, but for the black groups in samples 2 and 3 was somewhat low. KR20 figures need to be 0,7 or higher to be considered acceptable.

TABLE 5.1 MEANS, STANDARD DEVIATIONS AND KR20 STATISTICS FOR THE INTERMEDIATE MENTAL ALERTNESS TEST

SAMPLE 1

	ASIAN	BLACK	COLOURED	WHITE
N	206	207	102	99
mean	18,29	12,75	18,25	19,62
sd	4,92	4,36	4,97	4,58
KR20	0,80	0,75	0,80	0,77

SAMPLE 2

	BLACK
N	52
mean	12,19
sd	3,72
KR20	0,66

SAMPLE 3

	BLACK	COLOURED	WHITE
N	128	199	74
mean	12,32	16,33	15,15
sd	3,62	4,56	6,22
KR20	0,63	0,75	0,86

5.2.1.2 Black-white comparisons - Item bias results

Item bias analyses were undertaken on six black-white pairs.

1. Sample 1 blacks and whites
2. Sample 2 blacks and sample 3 whites

3. Sample 3 blacks and whites
4. Sample 2 blacks and sample 1 whites
5. Sample 3 blacks and sample 1 whites
6. Sample 1 blacks and sample 3 whites

The item bias detection procedure produced 10 items as biased in 50% or more of the comparisons (three or more comparisons). These were items 2, 4, 6, 15, 16, 20, 23, 25, 27, 28. From the baseline comparisons between samples from the same race (the black-black and white-white comparisons), items 2 and 4 emerged as biased in the white-white comparison, and were dropped from the set of items regarded as biased in the black-white comparison. Hence the final set of biased items included items 6 (alphabetic code), 15 (alphabetic reasoning), 16 (alphabetic code), 20 (verbal analogy), 23 (verbal reasoning), 25 (alphabetic series), 27 (alphabetic code) and 28 (alphabetic series). Of these eight items that emerged as biased, item 6 was biased in favour of blacks, while the other seven items were biased against blacks.

All the items in the test were categorised according to type. Of a total of 11 items based on the alphabet (alphabetic codes or alphabetic series), 6 of these emerged as biased items, i.e. six out of the eight biased items were based on the alphabet. Although there are no numeric codes in the test, there are 6 numeric series items and not one appeared biased. Thus there did seem to be a tendency for a certain type of item to be biased - namely the alphabetic codes and alphabetic series items. Item 6 was an alphabetic code, meaning that although many of the alphabetic items in the test were biased against blacks, one was biased in favour of blacks.

5.2.1.3 Coloured-white comparisons - Item bias results

Item bias procedures were applied to four pairs of data.

1. Sample 1 coloureds and whites
2. Sample 3 coloureds and whites

3. Sample 1 coloureds and sample 3 whites
4. Sample 3 coloureds and sample 1 whites

Three items emerged as biased from the item bias analyses between the coloured and white samples: items 2, 4, 20. However, items 2 and 4 also showed evidence of bias in the white-white and coloured-coloured baseline comparisons, hence only item 20 was taken as definitely biased. Item 20 is a verbal analogy and was biased against coloureds.

5.2.1.4 Asian-white comparisons - Item bias results

The available data enabled two Asian-white comparisons to be undertaken.

1. Sample 1 Asians and whites
2. Sample 1 Asians and sample 3 whites

No items emerged as biased in the Asian-white comparisons.

5.2.1.5 Summary of findings

The eight items identified as biased when all the comparisons were considered were deleted from the test and all the examinees' scores recalculated. The original and new means and standard deviations for the samples are reported in Table 5.2.

For sample 1, when the average black-white standard deviation was calculated and the difference between the black and white mean scores divided by this average standard deviation, the black-white mean difference on the old test was 1,46 standard deviation units. For the rescored version the difference was 1,24 standard deviation units. When the same calculations were performed for sample 3, the black-white mean score difference was 0,63 standard deviation units for the original test, and 0,33 standard deviation units for the rescored version. Thus, removing the biased items from the test did reduce the mean score difference

between blacks and whites slightly, although it did not altogether eliminate mean score differences.

For all of the samples completing the Intermediate Mental Alertness test, tables of item p values (difficulty values), item discrimination values (item-test total correlations) and the percentage of examinees not answering each item were computed for the different race groups. (The tables are presented in Appendix A. The biased items are highlighted.)

TABLE 5.2 ORIGINAL AND RESCORED MEANS AND STANDARD DEVIATIONS FOR THE INTERMEDIATE MENTAL ALERTNESS TEST

	MEAN		STANDARD DEVIATION	
	old	rescored *	old	rescored
SAMPLE 1				
Asians	18,29	14,01	4,92	3,38
blacks	12,75	10,54	4,36	3,38
coloureds	18,25	13,85	4,97	3,39
whites	19,62	14,68	4,58	3,17
SAMPLE 2				
blacks	12,19	10,29	3,72	2,94
SAMPLE 3				
Asians	17,14	12,86	3,48	2,91
blacks	12,32	10,23	3,62	2,99
coloureds	16,33	12,52	4,56	3,13
whites	15,15	11,36	6,22	4,59

* The means and standard deviations are lower for the rescored test because the old test is a 30 item test, whereas the rescored version comprises 22 items.

All the items were rank-ordered according to item difficulty (p-values) and item discriminability (item-total correlations) for the different groups and samples. Amongst the white samples, the biased items were of varying item difficulty. However, for the black, coloured and Asian groups there was a tendency for the biased items to be amongst those with the highest item difficulty values.

When the item-test total correlations were examined, it was found that for the white samples the biased items, although of varying discriminability, tended to be among the better discriminating items. This occurred in the Asian and coloured samples as well. However, for the black samples the biased items were amongst those items with the lowest item discrimination power. In particular, item 20, item 23, item 27 and item 28 (biased items) discriminated poorly amongst blacks, whereas they discriminated well in the other groups.

Many more blacks than whites did not complete the last few test items. Approximately one half of the black respondents did not answer item 30, the last test item, whereas approximately one quarter of the white respondents did not complete item 30. Research on testing in South Africa has suggested that blacks in particular usually have problems finishing tests in the allotted time period. However, the biased items were not those which most respondents did not complete, although there was a tendency for the biased items to be those towards the end of the test.

Overall, the results indicated that many items based on the alphabet, i.e. alphabetic codes and alphabetic series, were biased against blacks. Thus some adjustment should be made in the scores of black applicants for the bias in the alphabetic type items where this specific knowledge is not necessary to job performance. Furthermore, some consideration could be given to replacing item 20 (soil is related to the earth as oxygen is related to (?) - a) atmosphere b) nitrogen c) breathe d) gas e) life), which appeared to be biased against blacks and coloureds,

in a new version of the test.

The reader needs to bear in mind that these findings pertain to specific samples - samples of apprentice applicants. If item bias studies had been conducted on clerical job applicants a different pattern may have emerged. What one can say, however, is that for apprentice job applicants, several items do appear to be biased against blacks, and some correction should be made to counter the bias.

5.2.2 *HIGH LEVEL FIGURE CLASSIFICATION TEST*

5.2.2.1 Means, standard deviations and KR20 statistics

In Table 5.3 below, the means, standard deviations and KR20 figures for the samples that completed this test are reported.

When the mean scores for the samples were compared it could be seen that the results were similar to those of the Intermediate Mental Alertness. Generally whites obtained the highest mean scores, followed by Asians, coloureds and then blacks. Sample 3 was once again unusual in that the average score for the white group was below that of the coloured group. The standard deviations for the High Level Figure Classification Test also showed the same pattern as the Intermediate Mental Alertness. For all races, in all the samples, the internal consistency reliability figures were very good, 0,79 and higher.

5.2.2.2 Black-white comparisons - Item bias results

The same six black-white comparisons were undertaken as described in section 5.2.1.2. As a result of these analyses one item, item 17, was identified as biased. This item did not emerge as biased in the white-white and black-black baseline comparisons.

TABLE 5.3 MEANS, STANDARD DEVIATIONS AND KR20 STATISTICS FOR THE HIGH LEVEL FIGURE CLASSIFICATION TEST

SAMPLE 1

	ASIAN	BLACK	COLOURED	WHITE
N	206	207	71	77
mean	16,50	13,43	16,77	17,82
sd	4,55	5,44	4,35	4,09
KR20	0,84	0,87	0,82	0,80

SAMPLE 2

	BLACK
N	71
mean	12,49
sd	5,13
KR20	0,85

SAMPLE 3

	BLACK	COLOURED	WHITE
N	128	199	74
mean	13,22	16,75	15,14
sd	4,31	4,07	6,08
KR20	0,79	0,79	0,91

When the items of the High Level Figure Classification Test were examined, item 17 did appear to be different from the rest of the items in the test. Item 17 has a three-dimensional visual-perceptual component to it which the other test items do not have. Item 17 was biased against blacks.

5.2.2.3 Coloured-white comparisons - Item bias results

Item bias procedures were applied to the previously mentioned four coloured-white pairs (section 5.2.1.3). Only item 9 emerged

as biased more than 50% of the time in the four comparisons. This item was not biased in the baseline comparisons and was biased in favour of coloureds. It was difficult to determine a possible cause for this item appearing as biased. From an examination of the bias indices it could be ascertained that the amount of bias for item 9 was relatively minimal.

5.2.2.4 Asian-white comparisons - Item bias results

No items were biased in the two Asian-white item bias comparisons.

5.2.2.5 Summary of findings

When the results of all the bias analyses were combined, item 9 and item 17 were classified as biased. These two items were excluded from the test and the test rescored. The original and rescored means and standard deviations for the samples are reported in Table 5.4.

For the first sample, the average black-white standard deviation was computed and the difference between the black and white mean scores divided by this average standard deviation to produce a black-white mean difference on the original test of 0,95 standard deviation units. When the same calculations were performed for the rescored test, the black-white mean difference was exactly the same, i.e. 0,95 standard deviation units. The same procedure was followed for the third sample, producing a black-white mean score difference of 0,43 standard deviation units for the original test and 0,43 standard deviation units for the rescored test. Thus, removing the two biased items had no effect on the black-white mean score difference.

The tables containing the item difficulty (p-values) values, item-total correlations and percentage of examinees omitting each item are presented in Appendix B. Rank-ordering of the items according to difficulty and discriminability was undertaken.

TABLE 5.4 OLD AND RESCORED MEANS AND STANDARD DEVIATIONS FOR THE HIGH LEVEL FIGURE CLASSIFICATION TEST

	MEAN *		STANDARD DEVIATION	
	old	rescored	old	rescored
SAMPLE 1				
Asians	16,50	14,99	4,55	4,21
blacks	13,43	12,22	5,44	5,04
coloureds	16,77	15,30	4,35	3,89
whites	17,82	16,23	4,09	3,71
SAMPLE 2				
blacks	12,49	11,45	5,13	4,76
SAMPLE 3				
Asians	18,50	16,93	3,42	3,22
blacks	13,22	12,15	4,31	4,01
coloureds	16,75	15,26	4,07	3,82
whites	15,14	13,92	6,08	5,53

* The original test is a 24 item test. Hence, the original means and standard deviations are slightly higher than these same test statistics for the rescored 22 item version.

For the white samples, items 9 and 17 were of average item difficulty, and both items had virtually identical p-values. For Asians, blacks and coloureds, item 17 was one of the more difficult items and item 9 was one of the easiest items.

When the item-total correlations were examined, items 9 and 17 were amongst the best discriminating items for whites. For blacks, item 17 was among the poorest discriminating items, whereas item 9 was among the best. Compared to the other items, items 9 and 17 were average in discriminating power for the Asian

sample. An examination of the coloured samples showed that items 9 and 17 were of varying discriminability and no clear pattern emerged. Overall, item 9 discriminated very well in all groups and item 17 did not discriminate well in the black groups although high item-total correlations were obtained with the other groups.

Items 9 and 17 were not characterised by a high percentage of missing responses. Hence, bias did not appear to be due to examinees omitting these items. Once again, fairly large numbers of black and coloured examinees did not finish the test within the time limit. For example, thirty-eight percent of blacks from sample 2 did not complete item 24. Almost all the white examinees answered all test items.

In general, it appears that because item 17 seemed to function differently for the different groups and was clearly a different type of item to the remaining items in the High Level Figure Classification test, this item should be removed from the test. Ideally, because both item 9 and 17 were biased they should both be removed from the test and replaced with other unbiased items. Nevertheless, the amount of bias in the High Level Figure Classification test is small and the test can probably be used in its present form.

The HL FCT is one of the tests most often recommended for use with blacks. This is largely because of its nonverbal nature and the fact that during its construction specific attempts were made to standardise it on all race groups. In contrast, many of the earlier tests were standardised on, and norms developed for, whites only. Furthermore, black-white mean differences on the HL FCT are usually smaller than those observed for other tests used in industrial selection and the internal consistency reliabilities for all groups are high.

5.2.3 *BLOX TEST*

5.2.3.1 Means, standard deviations and KR20 statistics

In Table 5.5 a summary of the means, standard deviations and KR20 statistics for the subjects who completed the Blox test (A/80) is presented.

The pattern of means and standard deviations is similar to that obtained for the two tests already discussed. Generally, whites have the highest mean scores, followed by coloureds and Asians, then blacks. All internal consistency reliability figures are very good.

TABLE 5.5 MEANS, STANDARD DEVIATIONS AND KR20 STATISTICS FOR THE BLOX TEST

SAMPLE 1

	Asians	Blacks	Coloureds	Whites
N	206	207	102	99
mean	27,33	23,12	27,80	32,70
sd	7,04	7,05	6,94	4,98
KR20	0,85	0,83	0,85	0,76

SAMPLE 2

	Blacks
N	70
mean	24,13
sd	6,80
KR20	0,82

SAMPLE 3

	Blacks	Coloureds	Whites
N	128	200	74
mean	23,87	28,93	30,55
sd	6,76	6,30	7,58
KR20	0,82	0,82	0,88

5.2.3.2 Black-white comparisons - Item bias results

Three items showed up as biased in the analyses conducted on the six black-white pairs, but unbiased in the baseline studies. These were items 11, 12 and 24. It could not be ascertained why these items emerged as biased and others not. These three items were all biased against blacks.

5.2.3.3 Coloured-white comparisons - Item bias results

Three items appeared as biased in the coloured-white comparisons and not in the baseline analyses. These were items 21, 33 and 35. These three items were biased against coloureds. Once again no explanation could be found for the bias.

5.2.3.4 Asian-white comparisons - Item bias results

No biased items emerged from the analyses.

5.2.3.5 Summary of findings

When the results from all the analyses were combined, items 11, 12, 21, 24, 33 and 35 were biased. These items were omitted from the test and the test total scores recalculated for all examinees. The original and corrected means and standard deviations for the samples are reported in Table 5.6.

Once again the black-white mean difference in standard deviation units was computed. The difference between the black and white mean scores was divided by the average standard deviation for both groups. For the original test the black-white mean difference was 1,47 standard deviation units for sample 1. The black-white mean difference for the rescored test was 1,36 standard deviation units. When the same calculation was performed for the third sample, the black-white mean score difference was 1,01 standard deviation units for the original test and 0,89 standard deviation units for the rescored test. Thus, rescoring

the test without the biased items did seem to have some slight effect in favour of blacks.

TABLE 5.6 ORIGINAL AND RESCORED MEANS AND STANDARD DEVIATIONS FOR THE BLOX TEST

	MEAN*		STANDARD DEVIATION	
	old	rescored	old	rescored
SAMPLE 1				
Asians	27,33	23,84	7,04	6,00
blacks	23,12	20,30	7,05	6,35
coloureds	27,80	24,25	6,94	6,08
whites	32,70	28,01	4,98	4,27
SAMPLE 2				
blacks	24,13	21,17	6,80	6,00
SAMPLE 3				
Asians	32,29	27,50	5,87	5,60
blacks	23,87	21,14	6,76	6,01
coloureds	28,93	25,42	6,30	5,34
whites	30,55	26,35	7,58	6,40

* The original statistics were calculated on a 45 item test. The corrected scores are for a 39 item test. Subjects completing the Blox test have to answer 51 items, but the first 6 items are practice items and are not included in the calculation of means, norms, reliabilities etc.

Because three of the biased items were biased against coloureds, the same calculations were performed for sample 1 and sample 3 coloureds and whites. Rescoring the test without the biased items did slightly reduce the coloured-white mean difference (sample 1: 0,82 vs 0,73; sample 3: 0,23 vs 0,16).

The tables containing the item difficulty values, item-total correlations and percentage of examinees omitting each item are presented in Appendix C. For each group, items were rank-ordered according to difficulty and discriminability.

For whites the biased items were of varying item difficulty. That is the biased items were not particularly hard or easy for this group. For Asians, blacks and coloureds the biased items tended to be among the more difficult items.

Examination of the item-total correlations revealed that for Asians and whites the biased items, although of varying discriminability, tended to be among the better discriminating items. For blacks and coloureds the biased items varied in item discriminability with some discriminating well and others poorly. Items 11 and 24 discriminated well amongst whites and poorly amongst blacks. In particular item 24 had some very low values, including a negative item-total correlation in the black samples. Item 24 also did not discriminate well in the Asian and coloured groups.

The biased items were spread throughout the test, but nevertheless tended to be predominantly among the earlier test items. Therefore it seemed unlikely that bias was due to examinees not completing the test. In any event, the biased items were not those items that most examinees omitted. Virtually every examinee from each group answered all the biased items. As usual, items towards the end of the test were more likely to be omitted by examinees, but the percentage of each group responding to the last few items of the Blox was much higher than for the previous two tests, particularly for the black and coloured groups. This finding supported suggestions that in practice the Blox test is one of the few tests blacks find relatively easy to complete in the allotted time. Other groups also seemed to find it easier to complete the Blox.

5.2.4 *MECHANICAL COMPREHENSION TEST*

Fewer samples were obtained for this test than for the other three tests already discussed. This was not considered to be a problem because experience with the Mechanical Comprehension test had revealed that the test is not appropriate for use with black groups. A sample of Asian, black, coloured and white apprentice applicants in the sugar industry (sample 1) and a sample of black apprentice applicants tested by a development company in Natal (sample 2) was used in these comparisons.

5.2.4.1 Means, standard deviations and KR20 statistics

In Table 5.7 the means, standard deviations and KR20 figures for these samples are summarised.

As expected, the white mean was the highest, followed by the coloured, Asian and black mean scores. Standard deviations varied, but were fairly similar. The reliabilities were not good for any of the groups except perhaps the white sample. The very low reliability figures for the Asian and black groups in sample 1, despite the fairly large sample sizes, indicated problems with the use of this test for these groups. Similar, unacceptably low, reliability statistics have frequently been found for black groups on the Mechanical Comprehension test.

5.2.4.2 General

It has been recognised for some time that the Mechanical Comprehension test does not work well in the black population. Low reliabilities are very often reported. If a test has a low reliability it is impossible for it to attain reasonable validity as reliability affects the maximum validity possible. The problems experienced with this test, and the need for a mechanical test, has resulted in the construction of a new mechanical test battery.

TABLE 5.7 MEANS, STANDARD DEVIATIONS AND KR20 STATISTICS FOR THE MECHANICAL COMPREHENSION TEST

SAMPLE 1

	Asians	Blacks	Coloureds	Whites
N	206	207	102	99
mean	18,62	15,61	18,91	22,64
sd	4,29	4,15	4,59	5,33
KR20	0,53	0,52	0,60	0,71

SAMPLE 2

	Blacks
N	50
mean	14,44
sd	3,77
KR20	0,41

Item bias analyses were conducted on the samples and many items emerged as biased - more so than with the other tests. It is recommended that the Mechanical Comprehension test not be used and consequently no further analyses were undertaken on this test.

6. PREDICTIVE BIAS ANALYSES

6.1 INTRODUCTION

In practice, tests are used to predict some criterion measure, usually job or training performance.

Predictive bias research is concerned with the investigation of tests as predictors of criterion measures for different groups. According to Taylor (1987) the universally accepted definition of predictive bias is as follows: "A test is a biased predictor if there is a statistically significant difference between the major and minor groups in the slopes, or in the intercepts, or in the standard error of estimates of the regression lines of the two groups, when these regression parameters are derived from the estimated true scores of the persons within each group" (p. 17).

When we wish to predict a criterion score from a test score, a regression line can be computed which illustrates the linear relationship between test scores and criterion scores. According to the above definition, if a test is an unbiased predictor, the regression lines of test on criterion must be the same for each group.

Predictive bias research differs from item bias research in that some measure of job or training course success (a criterion measure) is necessary. Consequently, when predictive bias research is undertaken, the main problem is usually to find an appropriate criterion. A suitable criterion is one that is an unbiased measure of performance and is the same for the groups being compared.

A suitable criterion measure is not often readily available in companies. Because criterion measures differ across occupations and organisations, it is very difficult to generalise from one, or even from a few, predictive bias studies. For this reason, in

his report "Test bias: The roles and responsibilities of test user and test publisher", Taylor (1987) argued that the examination of tests for predictive bias was primarily the responsibility of test users.

In this chapter the results of some predictive bias research undertaken is reported. It was intended that some thorough predictive bias studies be conducted and presented as examples. However, there were many problems with the data and consequently the findings reported in this chapter should be regarded as tentative.

There has been very little predictive bias research conducted in South Africa to date. The few blacks employed in apprentice positions is one factor hampering this type of research, as is the criterion problem. It is often extremely difficult to obtain suitable criterion data, as was the case in this research. Many organisations either do not have this information or have information which varies widely from candidate to candidate. Furthermore, it is often extremely difficult to obtain large sample sizes for certain groups such as blacks, who are still sometimes poorly represented in apprentice positions. An added complication in this research was that many of the larger companies that had criterion data for reasonably large samples, either used different tests for apprentice selection to those investigated here, or else were not willing to allow the researchers access to their data.

Nevertheless, it was possible to obtain criterion data for the High Level Figure Classification test (HL FCT) and Blox test from one company that employed many apprentices. The criterion data available, in the form of two performance appraisal reports for each candidate, was excellent. In particular, one was a thorough measure of performance on apprentice training modules.

Test users cannot conclude from a single study that the same results apply to their apprentices. Although a test can be said

to have biased items, a test cannot be predictively biased per se. A test can only be predictively biased with respect to a specific criterion, and for this reason predictive bias in tests should be investigated in each individual company.

Despite the limitations of these predictive bias studies, this research can be used to illustrate to test users how to conduct a predictive bias study and subsequently how to deal with any bias that may emerge. After a great deal of predictive bias research has been performed, it may be possible to draw some clear conclusions with regard to specific tests and specific criterion measures.

6.2 TEST INFORMATION

The four tests most frequently used for apprentice selection were chosen to be examined for bias. Because the internal consistency reliability figures clearly indicated that the Mechanical Comprehension test is inappropriate for multicultural apprentice selection, it was excluded from the analysis.

The company from which the data was obtained had only been using the Intermediate Mental Alertness test for the past two years. Hence, insufficient information was available for this test and predictive bias analyses were only performed for the Blox and HL FCT. Descriptions of these tests can be found earlier in this report in section 3.1. At this point it is worth noting that these two tests demonstrated less item bias than the Intermediate Mental Alertness test, and the HL FCT in particular displayed little evidence of item bias.

6.3 CRITERION DATA

The criterion data used in a predictive bias study should be free of bias and equivalent for all the groups being compared. Because many of the different race groups attend separate technical

colleges, technical college results are often not suitable. Therefore, the author attempted to obtain criterion data in the form of direct on-the-job performance measures.

Two types of criterion data were made available to the author, together with the test scores, and used in the predictive bias research.

6.3.1 *Progress Evaluation Report (C1)*

The first criterion was obtained from a monthly progress evaluation report developed by the company. The apprentice's supervisor was required to rate the candidate on a scale of 1 to 10 on several job dimensions including factors such as initiative, efficiency, job quality, safety and interpersonal functioning. The applicant's rating for each factor was summed to produce an overall total score. A single percentage score (C1) was obtained by averaging all the total scores for the first year of apprenticeship.

6.3.2 *Efficiency Report (C2)*

An efficiency report developed by the company was completed for each apprentice. This report contained the percentages assigned to each apprentice after completion of sections of modules. This efficiency report is a very good, and relatively objective, performance appraisal measure in that it reflects the apprentice's on-the-job performance. C2 is a better criterion measure than C1 because it is more directly tied to the apprentice's performance on training modules and, hence, work performance. The second criterion measure (C2) was obtained by averaging the percentages assigned to each apprentice throughout their first year of apprenticeship.

6.4 DESCRIPTION OF THE SAMPLE

The company that supplied the information for the item bias

studies designated as sample 3 in section 3.2.3 was the same organisation that supplied the criterion data. All the subjects in this sample had been accepted as apprentices at a large motor company in the Eastern Cape.

In predictive studies it is required that test and criterion data are available for each apprentice. Therefore, each person who is included in the sample must have been tested and selected and have remained in the service of the company for one year. In this particular company most of those tested each year were not accepted (most of this data having been collected while a recession was in progress) and many of those accepted were not tested. A large percentage of the applicants accepted had an N3 and in accordance with company policy were not tested before being assigned a place. Thus, obtaining reasonable sample sizes proved difficult, particularly for the black apprentices, as few black people applied for apprentice positions and many of those who did had an N3 and were therefore not tested. After the performance appraisals and test scores of every single black apprentice in the history of the company had been obtained the final sample comprised only 12 cases. For the white (N=25) and coloured groups (N=53) data was obtained by collecting information for the past four years. It is important to note, however, that for each case the criterion information was based on the first year of apprenticeship, while the test scores were the results of testing undertaken when the apprentices were applicants.

The sample comprised blacks, coloureds and whites, all male. (The number of Asians employed as apprentices in this company was far too small to enable analyses to be undertaken on this group.)

6.5 PROCEDURE

In order to perform a predictive bias study there need to be two samples and for each member of a sample there must be at least one test score and one criterion score.

Three groups were available for comparison and two comparisons were undertaken: black-white and coloured-white. Because there were two tests and two criteria and each test could be used to predict a different criterion, eight bias analyses were possible and were performed as outlined in Table 6.1. The predictive bias analyses were carried out on the full 24 item HL FCT and 45 item Blox (i.e. none of the biased items were removed).

The predictive bias analyses were performed using a computer program written at the HSRC for this purpose. This computer program is available as part of the NIPR testing statistics package - NTS version 2 (Boeyens & Taylor, 1991). Because a test may be a biased predictor for a particular criterion if either the slopes, the intercepts or the standard error of estimates of the regression lines of test and criterion differ between the two samples, the program indicates whether the slopes, intercepts or standard error of estimates of the two groups are significantly different. Statistical tests are performed and the results of each test, as well as the significance level, is printed out for the three tests. Bias was considered to be present if a statistical test was significant at the 0,05 level. Because a separate statistical test is performed for each of the three parameters under investigation, bias was designated as present if at least one of the tests was significant.

TABLE 6.1 PREDICTIVE BIAS ANALYSES UNDERTAKEN

Predictor	Criterion	Pair
HL FCT	C1	Black-white
HL FCT	C1	Coloured-white
HL FCT	C2	Black-white
HL FCT	C2	Coloured-white
Blox	C1	Black-white
Blox	C1	Coloured-white
Blox	C2	Black-white
Blox	C2	Coloured-white

The program also allows the user to correct for the internal consistency reliability of the test for each sample. Because the company that supplied the apprentices' test data for the predictive bias studies also provided the apprentices' test scores for the item bias analyses in sample 3, the KR20 reliabilities for the HL FCT and Blox, calculated in the previous study, were used.

The results of the analyses are reported below. Only a brief overview is provided because the data available only enable tentative conclusions to be drawn.

6.6 PREDICTIVE BIAS RESULTS FOR THE HIGH LEVEL FIGURE CLASSIFICATION TEST

6.6.1 *High Level Figure Classification test and C1*

The predictive bias analyses between black and white samples showed evidence of predictive bias. A significant difference in the standard error of estimates between the two groups was obtained.

Predictive bias was present between the coloured and white samples. Once again, there was a significant difference in the standard error of estimates between the two groups.

6.6.2 *High Level Figure Classification test and C2*

When the analyses were run on the black and white samples, no significant differences emerged, indicating that predictive bias did not seem to be present.

However, when the coloured and white samples were compared, the standard error of estimates were once again significantly different.

Overall, it was standard error of estimate bias that was found for the High Level Figure Classification test.

6.7 PREDICTIVE BIAS RESULTS FOR THE BLOX TEST

6.7.1 *Blox test and C1*

The predictive bias analyses between the black and white group revealed a significant difference in slopes.

Between the white and coloured samples, bias was present as indicated by a significant difference in the standard error of estimates.

6.7.2 *Blox test and C2*

A significant difference between the intercepts was present when the black and white groups were compared.

As with many of the other findings, when the analyses were run on the coloured and white samples, a significant difference was found between the standard error of estimates.

6.8 SUMMARY

The results of the predictive bias studies suggest that some predictive bias may be present. Between white and coloured apprentices, bias occurred in the Blox test due to differences in the standard error of estimates for both criteria. The same finding occurred with respect to the High Level Figure Classification test. If it is borne in mind that standard error of estimate bias is often considered unimportant (Humphreys, 1986), it appears that there may possibly be no noteworthy predictive bias between the coloured and white apprentices for both tests.

The analyses performed on the Blox test showed signs of slope and intercept bias in the black-white comparison. Less black-white bias seemed evident for the High Level Figure Classification test. One of the two comparisons was biased, with standard error of estimate bias being present.

It must be emphasised that although predictive bias was found in this research, particularly with regard to the Blox test, users of these tests cannot conclude that these findings apply in their circumstances. A great number of predictive bias analyses, with more adequate data, are necessary before clear trends will become evident.

7. SUMMARY AND CONCLUSIONS

In the previous two chapters the results of two types of test bias analyses have been presented, namely item bias and predictive bias. These findings will now be discussed and suggestions for dealing with the bias presented. The results of the item bias studies will be discussed first. Several item bias studies were performed with relatively large samples sizes for all groups and tests, and consequently the results of the item bias research can be accepted with some degree of confidence as being representative of apprentice test performance.

7.1 ITEM BIAS

In general the white groups obtain the highest mean scores on the tests, followed by the Asian and coloured applicants, then the black subjects. However, this general finding is subject to differences among the applicants. If an organisation obtains applications from very well educated coloured and average white people, the mean scores for the coloured group may well be higher than those for the white group.

7.1.1 *Item Bias - Intermediate Mental Alertness test*

The Intermediate Mental Alertness is a test of general intelligence and is used fairly frequently in multicultural selection. Many applicants for technical and clerical jobs are required to complete this test.

A considerable amount of item bias was found when the black and white samples were compared. Eight out of the 30 items emerged as biased, with 7 of these being biased against the black apprentice applicants. Most of the biased items appeared to be alphabetic type items, i.e. alphabetic codes or alphabetic series. When these biased items were removed from the test and the test rescored, the black-white mean difference narrowed

somewhat, although a black-white mean difference still remained. Thus, removing the biased items did improve the test performance of the black examinees to some degree.

One item was biased against coloureds and no bias appeared when the Asian and white samples were compared.

For the black applicants the number of biased items was large, and some adjustment to their scores seems warranted if bias is to be eradicated from the selection procedure. In particular it was noted that success on the biased items did not seem to be essential to apprentice or artisan performance.

One way of dealing with item bias is to calculate the total scores for the black examinees for only the unbiased items. In this example the total scores will be calculated on 22 items. In order to compare these scores with those of applicants who have scores out of 30, the total scores for the black examinees are multiplied by 30 and divided by 22.

Two alternative methods for dealing with item bias are to make use of separate norms or bands (see section 7.3).

Further evidence was also available which suggested that the Intermediate Mental Alertness may not, in its present form, be completely appropriate for black applicants. The KR20 internal consistency reliability figures for the Asian, coloured and white groups were acceptable, but for the black group the figures were somewhat low.

A good test item is one that has a high item-total correlation. That is, it discriminates well between good and poor test performers. For the Asian, coloured and white examinees the biased items (almost all were biased against blacks) were among the better discriminating items, whereas for the black applicants the biased items were among the more poorly discriminating items. This finding is further evidence that the test items biased

against blacks are not suitable for the black examinees.

One final remark with regard to the test performance of black applicants is that, as is often reported in the literature, they found it more difficult than the other candidates to complete the test in the allotted time.

7.1.2 *Item bias - High Level Figure Classification test*

The HL FCT is a nonverbal test of general reasoning ability. Previous research and practical experience have shown this test to be one of the most suitable for all examinees. The KR20 internal consistency reliability figures for this test are usually high for all applicants, as was the case in this research.

On the whole the amount of bias emerging in this test was small. One item was biased against blacks and one item was biased in favour of coloureds. The item that was biased against the black apprentice applicants appeared to be a somewhat different item to the remaining items in the test. This item seemed to have a three-dimensional visual-perceptual component to it, unlike the remaining 23 test items.

Because of the small amount of bias in the test, removing the biased items did not alter the mean differences between the black and white groups at all. The item that was biased against blacks was a poorly discriminating item relative to the other items for the black applicants. Once again many blacks had problems completing the test within the time limit.

The small amount of item bias present in the HL FCT and the trivial effect it appeared to have on test performance implies that the test can probably be used in its present form for all groups. When a new test is developed, or an updated version produced, item bias procedures can be applied and new items substituted for the biased ones.

7.1.3 *Item bias - Blox*

The Blox test is a nonverbal test of spatial ability. As occurred with the other nonverbal test, the HL FCT, all the KR20 figures for all groups were acceptably high. Overall the amount of item bias was more than that found for the HL FCT but less than that for the Intermediate Mental Alertness (taking into account that the number of items differ for each test).

Three items were biased against blacks and three different items were biased against coloureds. The reason for the bias was not clear. When the 6 biased items were omitted from the test and the test rescored, the black-white mean difference decreased slightly, although a large difference still remained. The same finding occurred for the coloured-white mean difference. The same correction could be applied to the Blox test as was suggested for the Intermediate Mental Alertness. The test can be scored on the unbiased items and this score multiplied by 45 and divided by the number of unbiased items for each group. Separate norms or bands could also be used, but although separate norms may be one way of dealing with bias they do not provide an exact correction for bias.

The Blox test appeared to be one of the few tests that black examinees easily completed. In fact all groups appeared to find this test much easier to finish in the time allowed.

7.1.4 *Item bias - Mechanical Comprehension test*

Prior to the collection and examination of the data, it was suspected that this test was not suitable for use with black applicants. Despite the large number of employers using this test, many had complained about its unsuitability for black examinees. When the data collected for this research was analysed the KR20 figures were far too low to be acceptable, with the possible exception of the white sample. On this basis alone it can be recommended that the test should not be used for any

group, except perhaps the white applicants, and even then it is not recommended as the test is very old. The HSRC is presently developing a new Mechanical Battery which should be used in the future if employers wish to make use of a mechanical test.

7.1.5 *Item bias - Conclusion*

In this research item bias was conducted on tests already in use. This was because the issue of test bias has only recently become important in this country and the techniques to examine test bias are among the most recent psychometric methods to be developed. The correct procedure is to conduct item bias studies during the development of a test as part of the item analyses. Items which emerge as biased during trial runs of the test can then be replaced with items which do not appear to be biased.

Item bias is tied up with the meaning or construct validity of a test and thus, as Taylor has argued, an examination of item bias is largely the responsibility of the test constructor. The role of the test user would be to examine predictive bias. However, many of the tests that are on the market have not been examined for item bias with the result that many test users are concerned with item bias in the tests they administer[1]. If employers are to continue using tests then they must take note of item bias as it is not possible to immediately replace all tests with versions that have had the biased items removed. One point to bear in mind is that if test users perform item bias detection procedures it is preferable that they use a conditional method because the large mean differences between black and white samples can lead to problems with the unconditional item bias detection techniques.

When bias is found in a test this must be dealt with in some way if tests are to be properly used. It is <u>not appropriate</u> for test

---

[1]A few of the more recently developed HSRC tests on the market have been subjected to test bias procedures. Therefore test users, when purchasing tests, should confirm with the seller whether or not bias analyses have been conducted and for which groups.

users to alter tests (this is one of the reasons why item bias should be corrected by the test constructor before the test is marketed). However, some correction for the bias can be made in favour of examinees who are prejudiced by the bias. As was mentioned with regard to the Intermediate Mental Alertness and Blox, one way of correcting item bias is to calculate the scores for examinees using only the unbiased items, then multiply this by the total number of items in the test and divide by the number of unbiased items. However, before a correction is made for biased items, the items should be examined. If the items are judged to be highly related to the validity of the test with the result that the elimination of these items could substantially lower the test's validity, it may be best not to correct for the bias.

Some test users may not be in a position to conduct item bias studies. If the presence of bias is recognised in a case like this, separate norms may be used or the method of banding may be adopted (see section 7.3). If test users choose to use tests as they are in their present form, they need to be aware that the Blox and the Intermediate Mental Alertness seem to contain biased items for some applicants.

In addition to recognising bias and deciding on an appropriate way to deal with this, test users should look at other test information before deciding whether a test is suitable for a particular group of applicants. Sometimes test users apply tests that are clearly much too difficult for one or more groups. If a group is obtaining very low scores then the test is too difficult and should be replaced with another test. An examination of the internal consistency reliability figures should be performed. A test is unsuitable if low figures (for example less than 0,7) are observed.

Verbal tests have always been seen to be a problem when black applicants are tested because many of the applicants are not examined in their home language. The item bias analyses revealed

that the nonverbal tests did seem to have fewer biased items and had better KR20 figures for black apprentice applicants. For apprentice selection a high level of language ability is not crucial; hence, nonverbal tests can be useful.

Finally it must be noted that the item bias results of chapter 5 pertain to apprentice applicants. Item bias findings in the same tests may differ for applicants for other occupations. The number and type of biased items may also change over time as the environments and experiences of the different groups change.

7.2 PREDICTIVE BIAS

The predictive bias research reported in chapter 6 serves as an example of the type of analyses test users can conduct in their companies. Because sufficient data on the Intermediate Mental Alertness was unfortunately not available, predictive bias was examined for the Blox and HL FCT only.

Predictive bias was found to exist. Much of this was standard error of estimate bias and it was noted earlier that this type of bias is often considered trivial (slope and intercept bias being the more important types, Humphreys, 1986). In addition the presence of standard error of estimate bias in this research could have been due to sampling problems (uneven small samples).

If standard error of estimate bias is ignored, then no predictive bias emerged for the HL FCT between the black, coloured and white samples for both criteria, though predictive bias emerged between the black and white samples for the Blox test for both criteria. For the Blox test and C1 the predictive bias analyses between the black and white groups showed bias because of significant differences in slopes. Intercept bias was present for the Blox and C2. In both cases the bias was against the black applicants as the black apprentices had lower average test scores but higher average performance scores than the white group, particularly when intercept bias was present. For the Blox and C2 the black

group's average test score was 3 points below the white's, whereas the average performance score for the black apprentices was 4 percent higher.

These findings must be regarded as tentative, however, because the sample sizes were small and unequal for the different races. Large differences in sample sizes, particularly when samples are small, can sometimes result in sample standard deviations that differ widely from one another when the true population standard deviations may in fact be very similar. In such circumstances bias may arise spuriously due to sampling errors. Small samples may also preclude the attainment of significant bias results. Intercept bias, for example, is less readily detectable when samples are small. Small samples could make it more difficult to detect true bias (and could also be responsible for significant results when predictive bias is in fact absent). As more blacks are accepted into apprenticeships, data may be more readily available for further predictive bias studies to be performed.

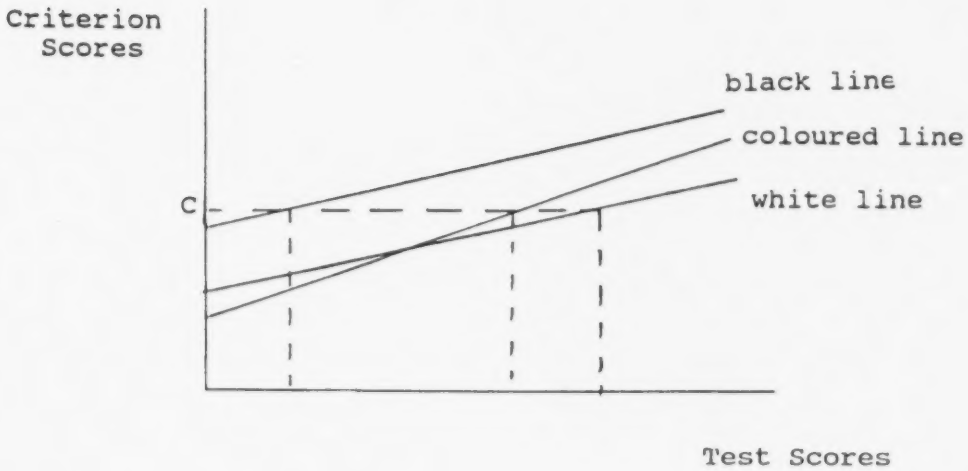
When predictive bias is present, the regression lines are significantly different for different groups. Using the same regression line or setting the same test score cut-offs when selecting will thus not overcome the bias.

This research showed some possible evidence of predictive bias for the Blox test between blacks and whites. The next concern is how to deal with the bias.

If the organisation has the resources to perform predictive bias analyses, a cut-off point on the criterion (i.e. job performance), above which performance is considered satisfactory, can be set. The regression lines of test on criterion can then be plotted for the various groups. A separate regression line should be produced for each group (see Figure 7.1 below). Test cut-off scores corresponding to the criterion cut-off score can subsequently be set for each group. If predictive bias is present, the cut-off scores on the tests are likely to differ for

the various groups.

Figure 7.1 ILLUSTRATION OF PREDICTIVE BIAS



It has become apparent that not all test users are in a position where they are able to conduct predictive bias studies and calculate regression lines. In this situation separate norms or bands may be used. These methods can also be used to deal with item bias.

7.3 DEALING WITH ITEM AND PREDICTIVE BIAS

In summary, it was noted in this research that one item on the HL FCT was biased against blacks, that three items of the Blox test were biased against the black group, that three Blox test items were biased against the coloured group and that 7 items of the Intermediate Mental Alertness test were biased against the black applicants. The predictive bias studies, though fraught with problems, revealed possible predictive bias against black applicants on the Blox test.

Clearly bias is present and should be dealt with. There are several ways in which to handle test bias.

1. Corrections to test scores

The organisation should consider if it has access to the resources (human and material) to conduct test bias research. If it has the item and predictive bias methods can be applied and the degree of bias in tests calculated. If bias is found the test scores can be corrected as suggested in this research.

1.1 If the test user has access to criterion data, then predictive bias studies can be performed. If the test user is able to perform predictive bias studies, the amount of predictive bias can be calculated for each group, regression lines computed and cut-offs set which may help to eliminate the predictive bias. It must be noted that the only way a company can ascertain if a test is having a negative effect on any particular group is by examining test performance and relating this to job performance. Therefore predictive bias studies should be a priority for test users.

1.2 If the test user is able to conduct item and predictive bias research, predictive bias analyses should be performed on the whole test first (biased items should not be removed). If predictive bias emerges, then item bias can be performed, any biased items deleted from the test and predictive bias analyses rerun on the test without the biased items. Any predictive bias that results can be corrected by examining the regression lines for the different groups. Regression lines of test on criterion can be plotted and different test cut-off scores set to help eliminate the bias.

1.3 If item bias research is the only approach taken and item bias is found, the test may be scored out of the number of

unbiased items. In order to equate these new scores with scores out of the original number of items the new score can be multiplied by the total number of items and divided by the number of unbiased items. For example, if there are 30 items in a test and 10 are biased against one group, the total score for the initial 20 unbiased items is first obtained. This score is then multiplied by 30 and divided by 20 to equate it with test scores out of 30 for other groups.

The above procedures represent methods for correcting bias once users have conducted bias studies. These are not the only ways in which to deal with bias. For some organisations it may not be possible to conduct bias studies. If this is the case, separate norms or bands may be used.

2 Separate Norms

It has been shown that item and possibly predictive bias can be expected for apprentice applicants. If companies do not have the resources to conduct bias studies, one way of dealing with test bias may be to make use of separate norms. In addition, if the organisation has a policy that the proportion of groups represented in the workforce is important, then making use of separate norms is one way to achieve this objective.

Although the use of separate norms has been criticised by many, the National Academy of Sciences in the USA has endorsed the use of separate norms for different race groups (Hartigan & Wigdor, 1989). They noted that when applicants are selected "top-down" according to test scores, the lower scoring group is adversely affected by this procedure. The reason for this is as follows.

When a test does not have perfect validity (and most have correlations with work performance of around 0,3 at the most) the applicants from the lower scoring group are more likely to be rejected when they could in fact have performed successfully on

the job. This is because when a test has low validity the average test scores differ between two groups more than does job performance i.e. there is greater overlap between groups on the job than in test performance. Individuals from the higher scoring group are more likely to do better on the test than on the job and benefit by being more likely to be selected when they may in fact fail. On the other hand, applicants from the lower scoring group are more likely to do better on the job than on the test and so are more likely to be rejected when they may have succeeded. Clearly selecting according to test scores, from the highest to the lowest, can be unfair to the lower scoring group.

If an employer decides to make use of separate norms, there is no need to make any further corrections for test bias i.e. to calculate appropriate cut-offs to offset the bias in the test or apply bands. Although the use of separate norms may seem to be a mechanism which greatly favours the lower scoring group, it must be noted that if the degree of bias in a test is large, the use of separate norms may be necessary to eradicate the effects of the bias.

Use of separate norms has been recommended when employers are concerned with the proportion of certain groups selected. In other words separate norms are often used as part of an affirmative action strategy. However, when bias is present in a test, particularly substantial bias, or a test has not been subjected to bias analyses, it is less easy to argue that the use of separate norms constitutes direct affirmative action procedures. If a test contains bias, then the use of separate norms may help to counter some of the bias. If an employer is concerned with the proportion of applicants from various groups selected into the company as well as with test bias, then the application of separate norms is a solution.

3 Bands

The use of bands has been recently suggested by some researchers

in the USA[2]. In order to apply this method, as with separate norms, no test bias studies need to have been undertaken. This approach, however, relative to separate norms, usually offers less advantage to the lower scoring group and may in fact not offset a large percentage of bias in a test.

No test is perfectly reliable as error is present in all test scores to some degree. Consequently it may not be possible to conclude that a score of 27 on a test is reliably different from a score of say 25. Test bands can be calculated so that all the scores within a particular band are considered to be not reliably different from one another. For example test scores of 20 to 25 may fall within one band. For the purposes of selection all these scores are regarded as the same.

The calculation of bands is fairly easy. The first step is determining the band width. This is a function of the standard deviation and reliability of the test and is obtained by multiplying several numbers together. Once the band width is known the highest test score obtained in a particular testing session is used to form the top of the band. If the band width is determined to be five points then the four points below the top score will also fall into the first band. Further bands falling below this top band can be calculated in a similar way. The above description applies to fixed bands; sliding bands may also be used.

The use of bands in selection can serve to increase to some extent the chances that an applicant from the lower scoring group will be accepted. For example if "top-down" selection is performed, applicants with test scores less than 23 might not be selected. However, if a band contains the scores 20-25, individuals with test scores of 21 or 22 may be accepted whereas

---

[2]Cascio and colleagues (personal communication) have recently proposed bands as a method of implementing fair selection. It has proved difficult to locate any literature on this topic which test users could easily obtain and for this reason parties interested in this method should contact the HSRC in Johannesburg.

some with scores of 23 or 24 may be turned away. Because all scores within a band are considered equal, if there are too many applicants falling within one band, alternative criteria for selection can come into play, for example increased representation of some groups in the workforce. One may decide to select all the black applicants and half the white applicants within the highest test score band.

Although the use of bands enables one to select more individuals from the lowest scoring group than "top-down" selection based on test scores, this need not be the approach adopted. The organisation is free to set their own criteria for selection within the bands.

7.4 CONCLUSION AND RECOMMENDATIONS

A survey identified that the four tests most frequently used for apprentice selection were the Intermediate Mental Alertness test, the Blox test, the High Level Figure Classification test and the Mechanical Comprehension test (Holburn, 1989). The test bias studies were performed on these tests.

Firstly it is recommended that the Mechanical Comprehension test should not be used for apprentice selection, particularly multicultural apprentice selection, as it is inappropriate for black, coloured and Asian samples. The HSRC has developed a new mechanical test battery to replace the old Mechanical Comprehension test.

With regard to the remaining three tests, the Intermediate Mental Alertness test was found to have 8 biased items out of a total of 30 items, the Blox 6 out of 45 items and the High Level Figure Classification test 2 out of 24 items. Predictive bias was also possibly present in the Blox test. Despite the presence of bias in these three tests, they are still considered to be among the best available measures for apprentice selection. The High Level Figure Classification test in particular is a highly suitable

test and greater weight should be given to these scores.

There are several ways in which employers can deal with bias in tests. The approach adopted by the employer will depend on the company's conception of fair selection. Two main approaches can be identified: selection on merit and some type of affirmative action programme aimed at increasing diversity in the workforce.

If a company decides to use its tests so that the only criterion is merit, then the bias techniques must be applied and the test scores of applicants altered to counter the bias. One problem with this approach, however, is that it is extremely difficult to measure all the bias in a test. There are other forms of bias besides item and predictive bias (e.g. pervasive bias, Taylor, 1987).

Approximately one third of the respondents to a survey on apprentice selection procedures (Holburn, 1989) indicated that they use separate norms for different race groups. This is one way to deal with the effects of bias. Although the bias may be overcome with this method, there is sometimes a degree of affirmative action applied when separate norms are used.

Another approach is to make use of test score bands and to select from within a band to increase diversity in the workplace. Although bands ensure that one is selecting individuals with high test scores, by selecting some individuals from the bottom of the band it may be possible for the goal of racial diversity in the workforce to be addressed as well. Bands are used in industry in South Africa, although to a very limited extent.

Another approach for an organisation may be to consider straight quotas. Regardless of the test scores of the different groups, a given percentage from each group is selected.

Although selectors may decide on the basis of the evidence presented in this report that tests should not be used for

apprentice selection because they contain bias, it must also be remembered that other selection instruments may also be biased. It is necessary that selectors bear in mind that when bias occurs in selection procedures, bias in tests is the easiest to ascertain and eradicate.

This document has been fairly technical as it is aimed at test users. For this reason many employees in an organisation may not understand some of the finer details of test bias. However, when tests are used they must, legally, be controlled by a psychologist, and an understanding of testing and test bias is expected of such persons. It is the responsibility of all parties to decide on the type of selection policy: Is selection to be based on merit? Is affirmative action the main priority? Or is it to be a combination of the two? The experts in charge of testing should be in a position to understand and perform testing in accordance with the company's selection policy. If the testing personnel require more information, an outside source such as the HSRC can be consulted.

REFERENCES

Anastasi, A (1985). Some emerging trends in psychological measurement: A fifty-year perspective. *Applied Psychological Measurement, 9(2)*, 121-138.

Angoff, W H (1982). Use of difficulty and discrimination indices for detecting item bias. In R A Berk (Ed.) *Handbook of methods for detecting test bias* (pp 96-116). Baltimore, MD: John Hopkins University Press.

Boeyens, J & Taylor, T R (1991). *Manual for the NIPR testing statistics package, version 2.0 k7.93*. Pretoria: Human Sciences Research Council.

Claassen, N C W (1990). *Die meting van intelligense in verskillende groepe met die Algemene Skolastiese Aanlegtoets (ASAT)*. Verslag P-118. Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Claassen, N C W & Cudeck, R (1985). The factor structure of the New South African Goup Test (NSAGT) in various population groups. *South African Journal of Psychology, 15(1)*, 1-10.

Cudeck, R & Claassen, N C W (1983). Structural equivalence of an intelligence test for two language groups. *South African Journal of Psychology, 13(1)*, 1-5.

Eells, K, Davis, A, Havighurst, R J, Herrick, V E & Tyler, R W (1951). *Intelligence and cultural differences: A study of cultural learning and problem-solving*. Chicago: The University of Chicago Press.

Halstead, M E & Du Toit, A (1983). *Blox test: Test administrator's manual (revised version) k7.66.* Johannesburg: National Institute for Personnel Research, Council for Scientific and Industrial Research.

Hartigan, J A & Wigdor, A K (Eds.) (1989). *Fairness in employment testing: Validity generalisation, minority issues, and the General Aptitude Test Battery.* Washington, D C: National Academy press.

Holburn, P T (1989). *Apprentice selection: An HSRC/NTB survey of policies and methods used in the with an emphasis on psychometric testing.* C/Pers 406. Pretoria: Human Sciences Research Council.

Humphreys, L G (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71(2),* 327-333.

Ironson, G H & Subkoviak, M J (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16,* 209-225.

Jensen, A R (1980). *Bias in mental testing.* London: Methuen.

Kaplan, R M (1985). The controversy related to the use of psychological tests. In B B Wolman (Ed.) *Handbook of intelligence: Theories, measurements and applications* (pp 465-504). New York: John Wiley & Sons.

Linn, R L, Levine, M V, Hastings, C N & Wardrop, J L (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159-173.

Lord, F M (1977). A study of item bias using item characteristic curve theory. In Y H Poortinga (Ed.) *Basic problems in cross-cultural psychology* (pp 19-29). Amsterdam: Swets & Zeitlinger.

Mellenbergh, G J (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

Owen, K (1986). *Toets- en itemsydigheid: Toepassing van die Senior Aanlegtoetse, Meganiese Insigtoets en Skolastiese Bekwaamheidsbattery op blanke, swart, kleurling- en Indiërtechnikonstudente*. Verslag P-66. Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Owen, K (1989). *Test and item bias: The suitability of the junior Aptitude Tests as a common test battery for white, Indian and black pupils in Standard 7*. Report P-96. Pretoria: Human Sciences Research Council.

Rudner, L M, Getson, P R & Knight, D L (1980). A monte carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17*, 1-10.

Shepard, L A, Camilli, G & Williams, D M (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

Shepard, L A, Camilli, G & Williams, D M (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Subkoviak, M J; Mack, J S; Ironson, G H & Craig, R D (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. *Journal of Educational Measurement, 21*, 49-58.

Taylor, T R (1987). *Test bias: The roles and responsibilities of test user and test publisher.* HSRC Special Report Pers 424. Pretoria: Human Sciences Research Council.

Van der Flier, H, Mellenbergh, G J, Ader, H J & Wijn, M (1984). An iterative item bias detection method. *Journal of Educational Measurement, 21,* 131-145.

Visser, B (1978). *Test administrator's manual k7.7 Mechanical Comprehension test (A/3/1).* Second edition revised. Pretoria: Human Sciences Research Council.

Werbeloff, M & Taylor, T R (1982). *Development and validation of the High Level Figure Classification Test.* Pers 338. Johannesburg: National Institute for Personnel Research, Council for Scientific and Industrial Research.

Werbeloff, M & Taylor, T R (1983). *Test administrator's manual for the High Level Figure Classification Test. k7.76.* Pretoria: Human Sciences Research Council.

Wilcocks, A. (1973). *Intermediate battery test administrator's manual (A/77). Revised Edition. k7.11.* Johannesburg: National Institute for Personnel Research, Council for Scientific and Industrial Research.

APPENDIX A

Intermediate Mental Alertness - Item difficulty values

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
1	0,98	0,86	0,94	0,94	0,93	0,95	0,97	0,82
2	0,84	0.57	0,58	0,52	0,88	0,26	0,88	0,30
3	0,88	0,82	0,87	0,79	0,87	0,85	0,95	0,81
4	0,84	0,78	0,81	0,81	0,90	0,58	0,90	0,47
5	0,68	0,53	0,54	0,59	0,72	0,61	0,71	0,57
6	**0,81**	**0,69**	**0,73**	**0,77**	**0,73**	**0,78**	**0,83**	**0,62**
7	0,79	0,68	0,65	0,64	0,77	0,74	0,73	0,69
8	0,91	0,71	0,65	0,73	0,84	0,81	0,84	0,76
9	0,87	0,79	0,73	0,73	0,86	0,85	0,93	0,74
10	0,81	0,53	0,52	0,52	0,80	0,69	0,84	0,65
11	0,73	0,63	0,62	0,69	0,66	0,76	0,77	0,65
12	0,62	0,57	0,52	0,47	0,73	0,67	0,74	0,62
13	0,53	0,20	0,29	0,27	0,47	0,57	0,48	0,43
14	0,56	0,41	0,40	0,37	0,62	0,55	0,63	0,54
15	**0,55**	**0,28**	**0,35**	**0,24**	**0,65**	**0,55**	**0,65**	**0,51**
16	**0,67**	**0,42**	**0,31**	**0,31**	**0,74**	**0,71**	**0,81**	**0,64**
17	0,80	0,55	0,44	0,43	0,74	0,72	0,77	0,58
18	0,68	0,47	0,33	0,30	0,75	0,66	0,77	0,61
19	0,82	0,66	0,65	0,70	0,75	0,83	0,79	0,69
20	**0,64**	**0,29**	**0,25**	**0,31**	**0,55**	**0,47**	**0,74**	**0,53**
21	0,50	0,31	0,23	0,24	0,36	0,34	0,47	0,41
22	0,52	0,27	0,23	0,23	0,49	0,47	0,60	0,31
23	**0,33**	**0,09**	**0,12**	**0,15**	**0,32**	**0,27**	**0,45**	**0,41**
24	0,24	0,07	0,10	0,09	0,25	0,16	0,38	0,19
25	**0,44**	**0,15**	**0,10**	**0,17**	**0,48**	**0,42**	**0,47**	**0,43**
26	0,10	0,06	0,08	0,07	0,19	0,23	0,19	0,22
27	**0,37**	**0,12**	**0,00**	**0,02**	**0,39**	**0,23**	**0,40**	**0,30**
28	**0,46**	**0,18**	**0,06**	**0,12**	**0,54**	**0,37**	**0,59**	**0,35**
29	0,14	0,06	0,08	0,07	0,09	0,09	0,24	0,14
0	0,17	0,03	0,04	0,03	0,19	0,14	0,11	0,18

A: Asians C: coloureds 1A = sample 1 Asians
B: blacks W: whites 3C = sample 3 coloureds
etc.

Intermediate Mental Alertness - Item-total correlations

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
1	0,34	0,31	0,39	0,08	0,09	0,11	0,01	0,39
2	0,39	0.33	0,19	0,37	0,28	0,15	0,25	0,26
3	0.24	0.34	0.24	0,28	0,18	0,29	0,06	0,38
4	0,18	0,32	0,03	0,30	0,10	0,26	0,10	0,37
5	0,32	0,39	0,25	0,42	0,39	0,26	0,33	0,59
6	**0,35**	**0,33**	**0,30**	**0,43**	**0,28**	**0,35**	**0,21**	**0,61**
7	0,41	0,34	0,42	0,15	0,27	0,23	0,47	0,56
8	0,28	0,25	0,19	0,43	0,33	0,32	0,28	0,49
9	0,38	0,28	0,20	0,46	0,35	0,43	0,34	0,57
10	0,39	0,51	0,09	0,39	0,53	0,24	0,46	0,56
11	0,39	0,31	0,10	0,18	0,32	0,17	0,30	0,60
12	0,34	0,36	0,37	0,18	0,26	0,36	0,25	0,47
13	0,19	0,18	0,08	0,15	0,11	0,17	0,06	0,33
14	0,40	0,17	0,13	0,15	0,27	0,35	0,26	0,35
15	**0,45**	**0,34**	**0,08**	**0,12**	**0,28**	**0,48**	**0,37**	**0,48**
16	**0,62**	**0,53**	**0,63**	**0,37**	**0,42**	**0,37**	**0,34**	**0,51**
17	0,50	0,38	0,32	0,31	0,46	0,26	0,51	0,50
18	0,52	0,51	0,36	0,45	0,51	0,43	0,26	0,54
19	0,35	0,19	0,41	0,58	0,40	0,26	0,35	0,40
20	**0,31**	**0,05**	**0,16**	**0,10**	**0,30**	**0,41**	**0,28**	**0,46**
21	0,09	0,23	0,23	-0,04	0,16	0,22	0,22	0,28
22	0,53	0,36	0,54	0,30	0,50	0,46	0,50	0,49
23	**0,12**	**-0,07**	**0,20**	**0,06**	**0,50**	**0,31**	**0,42**	**0,13**
24	0,33	0,10	0,24	-0,14	0,41	0,27	0,39	0,43
25	**0,46**	**0,40**	**0,61**	**0,04**	**0,51**	**0,43**	**0,47**	**0,45**
26	0,13	0,12	0,38	-0,00	0,44	0,23	0,41	0,05
27	**0,45**	**0,38**	**-0,00**	**0,13**	**0,57**	**0,39**	**0,56**	**0,30**
28	**0,45**	**0,40**	**0,21**	**0,16**	**0,47**	**0,44**	**0,40**	**0,48**
29	0,11	0,17	0,29	0,13	0,25	0,18	0,33	0,33
30	0,34	0,16	-0,28	-0,10	0,06	0,13	0,19	0,04

Intermediate Mental Alertness - percentage not completing item

Item	Sample 1A	1B	2B	3B	1C	3C	1W	3W
1	0,49	2,40	1,92	0,78	0,00	0,50	0,00	4,05
2	1,46	6.25	5,77	3,13	1,96	0,00	0,00	2,70
3	0,49	1,92	0,00	0,78	0,00	0,50	0,00	1,35
4	1,46	2,40	0,00	0,00	0,00	1,51	1,01	2,70
5	0,49	4,33	1,92	0,78	5,88	1,51	0,00	1,35
6	**2,43**	**3,37**	**3,85**	**3,13**	**0,98**	**1,51**	**1,01**	**2,70**
7	3,88	2,40	5,77	3,13	3,92	0,00	2,02	4,05
8	0,49	2,40	1,92	0,78	0,00	0,50	2,02	1,35
9	2,43	1,92	3,85	0,78	1,96	1,51	2,02	4,05
10	1,46	7,21	9,62	3,13	0,98	0,50	1,01	2,70
11	9,22	3,37	11,54	2,34	9,80	5,53	2,02	4,05
12	6,80	5,77	5,77	6,25	4,90	5,03	0,00	4,05
13	1,46	6,25	0,00	0,78	0,98	0,50	1,01	1,35
14	4,37	5,29	1,92	2,34	3,92	0,50	1,01	1,35
15	**4,37**	**4,33**	**0,00**	**2,34**	**0,98**	**0,00**	**2,02**	**1,35**
16	**11,17**	**18,75**	**23,08**	**17,97**	**9,80**	**5,03**	**4,04**	**4,05**
17	3,88	7,69	9,62	7,03	6,86	2,01	4,04	1,35
18	10,68	16,83	26,92	22,66	10,78	10,55	4,04	5,41
19	2,43	10,58	9,62	7,81	3,92	2,51	1,01	4,05
20	**1,46**	**7,21**	**5,77**	**6,25**	**1,96**	**1,01**	**1,01**	**2,70**
21	2,91	12,98	13,46	7,81	3,92	1,51	2,02	5,41
22	7,77	15,38	19,23	12,50	8,82	6,03	1,01	10,81
23	**12,62**	**22,12**	**25,00**	**17,19**	**15,69**	**10,05**	**4,04**	**9,46**
24	9,71	28,85	38,46	19,53	13,73	10,55	5,05	12,16
25	**16,99**	**33,65**	**36,54**	**28,13**	**20,59**	**14,57**	**11,11**	**12,16**
26	25,24	49,52	50,00	35,16	25,49	23,62	24,24	22,97
27	**20,39**	**49,04**	**57,69**	**46,09**	**22,55**	**29,15**	**20,20**	**25,68**
28	**24,27**	**52,88**	**50,00**	**43,75**	**27,45**	**29,15**	**20,20**	**24,32**
29	36,89	62,02	48,08	44,53	44,12	34,67	30,30	28,38
30	30,10	57,21	51,92	46,09	30,39	29,15	27,27	24,32

APPENDIX B

High Level Figure Classification Test - Item difficulty values

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
1	0,90	0,76	0,61	0,73	0,86	0,89	0,94	0,91
2	0,91	0,81	0,80	0,84	0,96	0,91	0,95	0,84
3	0,92	0,82	0,85	0,86	0,93	0,94	0,94	0,84
4	0,95	0,87	0,83	0,89	0,96	0,98	0,92	0,84
5	0,85	0,73	0,79	0,81	0,87	0,90	0,86	0,82
6	0,79	0,69	0,68	0,74	0,80	0,78	0,77	0,72
7	0,87	0,83	0,76	0,83	0,93	0,89	0,87	0,82
8	0,83	0,67	0,61	0,70	0,80	0,84	0,86	0,70
9	**0,89**	**0,78**	**0,70**	**0,80**	**0,86**	**0,90**	**0,79**	**0,72**
10	0,72	0,59	0,63	0,73	0,80	0,75	0,78	0,68
11	0,77	0,63	0,66	0,61	0,77	0,79	0,77	0,43
12	0,43	0,37	0,39	0,41	0,52	0,44	0,45	0,64
13	0,69	0,57	0,45	0,52	0,76	0,75	0,81	0,65
14	0,81	0,62	0,45	0,58	0,80	0,73	0,81	0,50
15	0,68	0,51	0,39	0,51	0,63	0,66	0,69	0,73
16	0,66	0,45	0,49	0,50	0,72	0,70	0,62	0,50
17	**0,63**	**0,43**	**0,34**	**0,27**	**0,62**	**0,59**	**0,79**	**0,73**
18	0,63	0,46	0,42	0,38	0,66	0,67	0,77	0,59
19	0,59	0,52	0,48	0,45	0,65	0,65	0,71	0,57
20	0,54	0,34	0,37	0,32	0,51	0,54	0,71	0,47
21	0,41	0,32	0,25	0,23	0,37	0,44	0,57	0,46
22	0,37	0,23	0,21	0,18	0,30	0,41	0,49	0,41
23	0,59	0,43	0,28	0,30	0,56	0,52	0,74	0,53
24	0,07	0,02	0,04	0,01	0,13	0,06	0,22	0,07

High Level Figure Classification Test - Item-total correlations

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
1	0,30	0,54	0,50	0,36	0,45	0,32	0,20	0,45
2	0,48	0,53	0,40	0,49	0,12	0,37	0,12	0,54
3	0,44	0,59	0,53	0,59	0,53	0,39	0,43	0,61
4	0,28	0,41	0,31	0,43	0,16	0,16	0,06	0,62
5	0,29	0,49	0,20	0,36	0,28	0,24	0,30	0,54
6	0,50	0,55	0,54	0,50	0,60	0,41	0,18	0,60
7	0,55	0,53	0,48	0,53	0,32	0,37	0,42	0,72
8	0,56	0,71	0,59	0,48	0,60	0,63	0,41	0,80
9	**0,46**	**0,62**	**0,60**	**0,53**	**0,50**	**0,42**	**0,46**	**0,68**
10	0,40	0,49	0,44	0,42	0,32	0,29	0,28	0,58
11	0,29	0,36	0,25	0,10	0,32	0,19	0,40	0,49
12	0,23	0,31	0,19	0,24	0,35	0,12	0,31	0,26
13	0,30	0,40	0,34	0,29	0,38	0,41	0,54	0,53
14	0,60	0,60	0,58	0,48	0,54	0,57	0,51	0,59
15	0,43	0,47	0,44	0,22	0,52	0,35	0,37	0,53
16	0,38	0,41	0,47	0,18	0,33	0,39	0,47	0,33
17	**0,41**	**0,30**	**0,26**	**0,23**	**0,65**	**0,26**	**0,57**	**0,55**
18	0,58	0,51	0,61	0,39	0,41	0,58	0,42	0,64
19	0,48	0,46	0,56	0,54	0,35	0,36	0,40	0,47
20	0,45	0,51	0,54	0,37	0,37	0,48	0,53	0,49
21	0,53	0,46	0,40	0,34	0,54	0,48	0,46	0,63
22	0,56	0,36	0,46	0,31	0,38	0,49	0,32	0,55
23	0,49	0,56	0,57	0,49	0,39	0,47	0,47	0,60
24	0,12	0,05	0,03	0,12	0,20	0,21	0,30	0,27

High Level Figure Classification Test - percentage not completing item

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
1	0,97	4,33	7,04	6,25	5,63	2,51	0,00	1,35
2	0,49	0,48	4,23	0,78	0,00	1,01	0,00	0,00
3	1,94	0,96	4,23	2,34	0,00	0,50	0,00	2,70
4	0,49	0,96	0,00	0,00	0,00	0,00	0,00	0,00
5	0,97	4,33	4,23	0,00	1,41	2,51	0,00	1,35
6	1,94	3,85	0,00	2,34	0,00	1,01	0,00	1,35
7	0,97	0,48	4,23	0,78	2,82	1,01	0,00	1,35
8	0,49	1,92	2,82	0,00	1,41	2,01	0,00	0,00
9	**0,49**	**1,44**	**9,86**	**2,34**	**2,82**	**1,01**	**0,00**	**2,70**
10	8,25	8,17	8,45	6,25	11,27	6,03	5,19	2,70
11	1,94	2,88	0,00	5,47	1,41	2,01	0,00	2,70
12	3,88	6,25	2,82	1,56	4,23	4,02	2,60	2,70
13	2,91	7,69	9,86	4,69	4,23	3,02	0,00	1,35
14	2,91	6,73	11,27	5,47	7,04	4,02	3,90	0,00
15	3,40	9,62	12,68	5,47	5,63	5,53	1,30	1,35
16	1,46	6,73	14,08	8,59	7,04	4,02	2,60	0,00
17	**4,37**	**4,81**	**12,68**	**7,81**	**4,23**	**3,52**	**1,30**	**1,35**
18	1,46	6,25	14,08	7,81	2,82	4,02	1,30	1,35
19	2,43	6,25	15,49	8,59	4,23	2,51	3,90	1,35
20	6,80	18,75	19,72	14,84	8,45	3,52	5,19	2,70
21	3,40	12,50	21,13	16,41	11,27	8,04	5,19	1,35
22	7,77	21,63	28,17	21,09	15,49	17,59	5,19	2,70
23	4,85	17,79	30,99	25,78	15,49	13,57	5,19	2,70
24	9,71	24,52	38,03	28,13	12,68	16,58	5,19	4,05

APPENDIX C

Blox - Item difficulty values

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
7	0,88	0,77	0,79	0,92	0,90	0,91	0,98	0,88
8	0,91	0,90	0,91	0,91	0,95	0,98	0,98	0,96
9	0,68	0,52	0,46	0,58	0,70	0,62	0,86	0,77
10	0,96	0,94	0,91	0,96	1,00	0,98	1,00	0,97
11	**0,73**	**0,56**	**0,53**	**0,55**	**0,75**	**0,61**	**0,8[illegible]**	**0,78**
12	**0,60**	**0,53**	**0,56**	**0,49**	**0,66**	**0,66**	**0,84**	**0,81**
13	0,49	0,47	0,46	0,41	0,48	0,45	0,67	0,61
14	0,80	0,71	0,80	0,66	0,76	0,82	0,91	0,81
15	0,37	0,36	0,43	0,31	0,43	0,42	0,48	0,41
16	0,84	0,74	0,76	0,79	0,83	0,87	0,98	0,92
17	0,82	0,64	0,66	0,73	0,86	0,88	0,97	0,82
18	0,82	0,78	0,80	0,75	0,83	0,86	0,94	0,89
19	0,92	0,83	0,90	0,85	0,90	0,91	0,95	0,86
20	0,77	0,62	0,67	0,64	0,77	0,77	0,88	0,78
21	**0,60**	**0,54**	**0,54**	**0,46**	**0,65**	**0,67**	**0,87**	**0,77**
22	0,58	0,46	0,47	0,47	0,58	0,64	0,82	0,69
23	0,60	0,56	0,46	0,62	0,63	0,68	0,73	0,77
24	**0,45**	**0,28**	**0,37**	**0,24**	**0,35**	**0,42**	**0,57**	**0,51**
25	0,91	0,86	0,94	0,90	0,91	0,95	0,93	0,91
26	0,84	0,72	0,70	0,78	0,90	0,94	0,95	0,88
27	0,80	0,60	0,61	0,73	0,77	0,89	0,95	0,89
28	0,82	0,63	0,64	0,68	0,81	0,89	0,94	0,91
29	0,74	0,51	0,54	0,60	0,76	0,81	0,89	0,80
30	0,72	0,58	0,63	0,60	0,75	0,78	0,88	0,80
31	0,72	0,56	0,59	0,68	0,75	0,79	0,76	0,80
32	0,80	0,64	0,73	0,63	0,74	0,82	0,94	0,84
33	**0,73**	**0,57**	**0,53**	**0,61**	**0,65**	**0,65**	**0,88**	**0,76**
34	0,70	0,52	0,71	0,58	0,69	0,67	0,82	0,72
35	**0,37**	**0,34**	**0,43**	**0,38**	**0,49**	**0,50**	**0,71**	**0,57**
36	0,68	0,58	0,69	0,63	0,75	0,77	0,80	0,77
37	0,30	0,22	0,26	0,20	0,35	0,28	0,42	0,34
38	0,25	0,29	0,34	0,28	0,35	0,37	0,48	0,34
39	0,38	0,31	0,33	0,42	0,42	0,39	0,46	0,50
40	0,36	0,23	0,41	0,28	0,27	0,44	0,36	0,43
41	0,53	0,36	0,40	0,49	0,48	0,53	0,63	0,65
42	0,45	0,36	0,37	0,38	0,39	0,52	0,66	0,54
43	0,69	0,62	0,63	0,59	0,70	0,78	0,85	0,77
44	0,64	0,59	0,56	0,45	0,68	0,75	0,79	0,74
45	0,60	0,43	0,37	0,43	0,66	0,65	0,75	0,65
46	0,21	0,21	0,11	0,18	0,15	0,19	0,25	0,27
47	0,23	0,25	0,23	0,21	0,26	0,25	0,19	0,28
48	0,30	0,26	0,29	0,24	0,32	0,33	0,33	0,38
49	0,31	0,29	0,21	0,24	0,27	0,27	0,32	0,35
50	0,21	0,19	0,21	0,16	0,20	0,22	0,27	0,39
51	0,20	0,19	0,19	0,20	0,27	0,26	0,24	0,27

Blox - Item-total correlations

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
7	0,26	0,28	0,17	0,15	0,24	0,27	-0,08	0,45
8	0,30	0,45	0,18	0,22	0,23	0,11	-0,04	0,14
9	0,42	0,49	0,31	0,49	0,60	0,27	0,41	0,48
10	0,37	0,41	0,21	0,19	-0,00	0,34	-0,00	0,42
11	**0,47**	**0,24**	**0,36**	**0,12**	**0,41**	**0,41**	**0,46**	**0,52**
12	**0,31**	**0,16**	**0,31**	**0,31**	**0,12**	**0,30**	**0,35**	**0,30**
13	0,40	0,34	0,14	0,34	0,30	0,24	0,26	0,54
14	0,22	0,41	0,23	0,40	0,32	0,24	0,17	0,50
15	-0,04	0,09	0,12	0,12	0,10	0,13	0,02	0,13
16	0,51	0,57	0,66	0,46	0,64	0,48	0,24	0,53
17	0,54	0,61	0,59	0,54	0,63	0,56	0,07	0,70
18	0,31	0,44	0,37	0,33	0,48	0,44	0,27	0,41
19	0,32	0,45	0,34	0,28	0,41	0,42	0,26	0,46
20	0,31	0,38	0,33	0,39	0,45	0,27	0,17	0,35
21	**0,51**	**0,35**	**0,41**	**0,36**	**0,40**	**0,44**	**0,36**	**0,62**
22	0,48	0,46	0,41	0,25	0,45	0,41	0,43	0,50
23	0,35	0,15	0,27	0,07	0,34	-0,01	0,17	-0,02
24	**0,24**	**0,02**	**-0,05**	**0,10**	**0,28**	**0,25**	**0,34**	**0,50**
25	0,37	0,18	0,19	0,20	0,17	0,31	0,15	0,04
26	0,64	0,54	0,53	0,53	0,49	0,41	0,24	0,43
27	0,64	0,59	0,49	0,50	0,57	0,45	0,30	0,36
28	0,52	0,65	0,50	0,60	0,63	0,60	0,38	0,58
29	0,55	0,48	0,49	0,49	0,51	0,49	0,28	0,57
30	0,48	0,43	0,47	0,52	0,41	0,45	0,32	0,54
31	0,42	0,37	0,30	0,30	0,39	0,33	0,41	0,38
32	0,55	0,61	0,56	0,45	0,71	0,62	0,31	0,57
33	**0,48**	**0,38**	**0,30**	**0,51**	**0,44**	**0,25**	**0,17**	**0,46**
34	0,36	0,38	0,39	0,47	0,56	0,34	0,15	0,27
35	**0,33**	**0,23**	**0,41**	**0,18**	**0,24**	**0,37**	**0,23**	**0,37**
36	0,50	0,46	0,40	0,48	0,50	0,41	0,35	0,66
37	0,16	0,05	0,01	0,19	0,43	0,15	0,41	0,33
38	0,29	0,18	0,23	0,19	0,36	0,30	0,46	0,39
39	0,21	0,26	0,31	0,30	0,18	0,24	0,14	0,17
40	0,33	0,22	0,16	0,31	0,19	0,30	0,46	0,28
41	0,31	0,24	-0,01	0,24	0,13	0,32	0,41	0,42
42	0,33	0,31	0,40	0,22	0,25	0,35	0,52	0,33
43	0,26	0,23	0,36	0,50	0,32	0,27	0,11	0,25
44	0,32	0,39	0,31	0,25	0,22	0,27	0,10	0,43
45	0,39	0,27	0,35	0,19	0,27	0,36	0,02	0,20
46	0,05	-0,11	-0,08	0,10	0,10	0,17	0,22	0,22
47	-0,06	0,05	0,10	0,14	0,04	0,21	0,18	0,26
48	0,17	0,04	0,18	0,06	0,10	0,14	0,26	0,47
49	0,16	0,19	0,31	0,12	0,11	0,22	0,32	0,41
50	0,16	-0,02	0,26	0,15	0,14	0,19	0,18	0,28
51	0,05	0,05	0,18	0,21	0,22	0,18	0,36	0,31

Blox - percentage not completing item

	Sample							
Item	1A	1B	2B	3B	1C	3C	1W	3W
7	0,00	0,00	0,00	0,78	0,00	0,00	0,00	0,00
8	0,49	0,48	0,00	0,78	0,98	0,00	0,00	0,00
9	0,97	1,44	1,43	0,78	1,96	1,00	1,01	0,00
10	0,49	0,00	1,43	0,78	0,00	0,00	0,00	0,00
11	**1,46**	**0,96**	**0,00**	**0,00**	**0,98**	**1,50**	**0,00**	**0,00**
12	**4,85**	**1,44**	**0,00**	**3,13**	**3,92**	**1,50**	**2,02**	**2,70**
13	4,85	4,33	4,29	5,47	5,88	4,50	2,02	1,35
14	3,88	1,92	1,43	3,13	5,88	3,00	0,00	0,00
15	3,88	0,00	2,86	5,47	4,90	4,00	0,00	1,35
16	0,00	0,00	1,43	0,78	0,00	1,00	0,00	0,00
17	0,49	1,44	1,43	0,00	0,00	0,50	1,01	0,00
18	2,43	0,48	0,00	3,13	0,00	0,50	0,00	0,00
19	0,00	0,00	0,00	1,56	0,00	0,50	0,00	0,00
20	0,49	2,40	1,43	1,56	0,00	0,50	0,00	0,00
21	**1,46**	**0,00**	**2,86**	**0,78**	**0,98**	**2,00**	**0,00**	**0,00**
22	2,43	5,29	7,14	4,69	8,82	5,50	0,00	2,70
23	2,43	2,40	1,43	1,56	1,96	1,00	0,00	0,00
24	**2,43**	**1,92**	**1,43**	**3,13**	**1,96**	**1,50**	**2,02**	**1,35**
25	0,00	0,00	0,00	0,78	0,98	0,50	0,00	0,00
26	0,00	0,00	0,00	0,78	0,00	0,50	0,00	1,35
27	0,00	0,00	0,00	0,78	0,98	0,50	0,00	1,35
28	0,49	0,00	0,00	0,00	0,00	0,50	0,00	1,35
29	0,00	0,00	0,00	0,78	2,94	0,50	0,00	0,00
30	0,00	0,48	0,00	0,78	1,96	0,50	0,00	0,00
31	1,46	2,40	4,29	3,91	3,92	3,00	0,00	1,35
32	0,00	0,00	1,43	1,56	2,94	2,00	0,00	0,00
33	**0,00**	**0,48**	**2,86**	**1,56**	**4,90**	**2,00**	**0,00**	**1,35**
34	0,49	0,00	1,43	3,13	0,98	1,50	0,00	0,00
35	**2,43**	**1,92**	**4,29**	**5,47**	**2,94**	**2,50**	**0,00**	**1,35**
36	0,97	0,00	2,86	2,34	1,96	1,00	0,00	1,35
37	2,43	1,44	5,71	3,13	0,98	3,00	1,01	1,35
38	2,43	1,44	2,86	5,47	2,94	2,50	1,01	1,35
39	4,85	2,40	2,86	6,25	7,84	2,50	1,01	0,00
40	6,31	3,85	4,29	7,81	4,90	6,50	5,05	0,00
41	3,88	3,85	5,71	3,91	6,86	3,50	3,03	0,00
42	2,91	4,33	5,71	7,03	4,90	4,50	3,03	0,00
43	0,97	1,44	5,71	8,59	0,98	4,50	0,00	1,35
44	0,97	0,48	7,14	7,03	0,98	4,50	1,01	0,00
45	1,46	2,88	8,57	8,59	0,98	5,00	1,01	1,35
46	2,43	3,37	8,57	10,16	6,86	6,00	0,00	1,35
47	6,31	4,81	11,43	11,72	4,90	7,50	1,01	1,35
48	2,91	5,29	7,14	13,28	2,94	8,00	3,03	0,00
49	5,83	4,33	14,29	16,41	7,84	7,50	5,05	0,00
50	4,85	5,77	12,86	14,06	7,84	8,50	4,04	0,00
51	3,40	2,88	14,29	15,63	3,92	8,00	2,02	1,35

ISBN 0 7969 1254 8

BEST COPY AVAILABLE